UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MARCO AURÉLIO SCHÜNKE

Aplicação de Algoritmos de Classificação para Análise dos Fatores que Influenciam na Predição do Fator de Impacto em Redes Sociais

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Dante Augusto Couto Barone

Porto Alegre
2015

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Schünke, Marco Aurélio
Aplicação de Algoritmos de Classificação para Análise dos Fatores que Influenciam na Predição do Fator de Impacto em Redes Sociais / Marco Aurélio Schünke. – 2015.
15 f.:il.
Orientador: Dante Augusto Couto Barone.
Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2015.
1.Predição. 2.Mineração de Dados 3.Redes Sociais, Orientador. Dante Augusto Couto Barone. Aplicação de Algoritmos de Classificação para Análise dos Fatores que Influenciam na Predição do Fator de Impacto em Redes Sociais

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus.

Em seguida, especialmente aos meus pais por sempre terem me incentivado a estudar e nunca desistir dos nossos sonhos e também por todo o suporte prestado para que eu pudesse evoluir na minha formação pessoal, profissional e principalmente acadêmica no âmbito da pesquisa.

Agradeço também a todos meus professores e colegas de trabalho da UNISC – Universidade de Santa Cruz do Sul, em especial ao Setor de Informática da universidade cujos valores éticos e morais sempre foram incentivados.

Agradeço também às demais pessoas que estiveram comigo durante essa fase da minha vida, como meus colegas de trabalho do SISTEMA FIERGS / SENAI – Serviço Nacional de Aprendizagem Industrial e em especial a coordenação pelo apoio e incentivo para que eu escrevesse esta dissertação.

Agradeço aos meus colegas de pesquisa do SISTEMA FIERGS / FATEC / SENAI - Faculdade SENAI de Tecnologia aos professores e colegas do Projeto SIGA-i o qual tenho o privilégio de fazer parte como bolsista pesquisador IEL/CNPq INOVA TALENTOS PROGRAMA RHAE TRAINEE CNPq/IEL do Instituto Euvaldo Lodi/Núcleo Central IEL/NC, em parceria com o Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq na Faculdade de Tecnologia do SENAI Porto Alegre - FATEC. RS. Brasil.

Também a todos meus professores da UFRGS que iluminaram os caminhos guiando-me no despertar da minha curiosidade pela Inteligência Artificial.

Além de meus colegas do grupo de pesquisa Web Science do Instituto de informática da Universidade Federal do Rio Grande do Sul que compartilharam o conhecimento de outras técnicas trabalhadas.

E em especial ao meu orientador Prof. Dr. Dante Augusto Couto Barone, pela dedicação, motivação e notório saber no campo da Web Science, além da orientação na realização desta dissertação.

RESUMO

Atualmente empresas como Google e Facebook fazem parte da lista das maiores companhias do mundo. O investimento em publicidade e criação de páginas para a divulgação de anúncios e marcas, tem levado o Facebook a uma posição de destaque neste cenário. Neste contexto, o presente trabalho tem o objetivo de analisar e predizer o número de interações em notícias divulgadas em cinco páginas de fãs, que se constituem nas mais acessadas da Rede Social Facebook no Brasil. Como contribuição propõem-se determinar o fator de impacto de publicações, considerando a média de três características mencionadas, o número de curtidas, o número de comentários e o número de vezes que a notícia foi compartilhada. Serão avaliados resultados da aplicação de diferentes técnicas para a classificação, além da influência de características relacionadas a palavras e termos mais frequentes, verificando qual combinação produz melhores resultados no processo de gerar um modelo de aprendizado para prever o Fator de Impacto de notícias publicadas nas páginas de fãs da Rede Social Facebook. Apresenta-se também os motivos que podem exercer influência no fator de impacto através do processo de descoberta de conhecimentos em base de dados e também fazendo uso de técnicas de processamento de linguagem natural com o objetivo de atender a expectativa do trabalho.

Palavras-chave: Predição, Mineração de Dados, Redes Sociais.

This should be the title in English

ABSTRACT

Currently companies as Google and Facebook are on the top of the largest companies in the world and according to news released on the website tecmundo the main reason that led to this privileged position, in particular Facebook, appears to be the result of its investments in publicity focused on mobile devices through general advertisements in its own social network. In this context the present research aims to estimate the number of news interactions published on the five most accessed fans pages of Facebook Social Network in Brazil. Are considered examples of interactions in this study the number of likes, the number of comments and also the amount of times a message was shared. As also disclose attributes that influence interactions. As a contribution is proposed the impact factor of a publication, considering the average of three mentioned interactions, the number of likes, the number of comments and also the number of times the news was shared, in order to improve the results in predicting interactions of a fan page of Facebook Social Network.

In addition to analyze the results of prediction algorithms applying different techniques of text pre- processing checking which combination produces best results in generating a learning process model to foresee the impact of news published on Facebooks fan pages and exhibit reasons that may influence the impact factor through the discovering process of database knowledge, from the feeling analysis as well as making use of processing of natural language techniques in order to fulfill work expectation.

Keywords: Prediction, Data Mining, Social Networks.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
DCBD	Descoberta de Conhecimento em Base de Dados
DICIO	Dicionário Online de Português
FQL	Facebook Query Language
FI	Fator de impacto
CU	Número de Curtidas de uma mensagem
CO	Número de Comentários de uma mensagem
CM	Número de Compartilhamentos de uma mensagem
FA	Fator de impacto alto
FM	Fator de impacto médio
FB	Fator de impacto baixo
IA	Inteligência Artificial
VP	Verdadeiros Positivos
FP	Falsos Positivos
VN	Verdadeiros Negativos
FN	Falsos Negativos
MYSQL	My Structured Query Language
SQL	Structured Query Language
SGBD	Sistema de Gerenciamento de Base de Dados
KDD	Knowledge Discovery in Databases
NLTK	Natural Language Toolkit
PLN	Processamento da Linguagem Natural
WEKA	Waikato Environment for Knowledge Analysis
UFRGS	Universidade Federal do Rio Grande do Sul

SUMÁRIO

1 INTRODUÇÃO

No início, as Redes Sociais eram utilizadas apenas como um entretenimento para alguns usuários entusiastas e hoje são utilizados para muitos propósitos, como por exemplo, divulgar viagens, produtos, empresas, fotos, vídeos, links, histórias, eventos, promover a educação e ainda difundir notícias que podem conduzir importantes mudanças em nossa sociedade, conforme (BUZA, 2012) como as revoluções no mundo Islâmico, ou eleições presidenciais dos EUA.

Redes Sociais constituem-se, em uma nova forma de comunicação, socializando relações entre pessoas através da tecnologia, aproximando pessoas compartilhando pensamentos, ideias e sonhos em comum, vencendo obstáculos como a distância e como consequência acelerando o processo para disseminar informações. Além disso, apesar das incontáveis interações geradas nas redes sociais que contribuem para o aumento de informações de domínios diversos, torna-se importante identificar formas de revelar os fatores que influenciam as interações, neste trabalho são analisado os seguintes fatores, o número de palavras presentes em cada mensagem, o número de curtidas, o número de comentários, o número de compartilhamentos, o número de unigram, o número de bigram, o número de trigram, o número de Ngram, o número de mensagens por mês, o número de mensagens por dia da semana, o número de mensagens por turno e o número de mensagens por tipos. E com isso é relevante realizar uma comparação entre as técnicas de predição existentes analisando sua eficácia no sentido de identificar os resultados sobre a possibilidade de prever o impacto de uma notícia, utilizando técnicas de Inteligência Artificial para obter as respostas na identificação de padrões.

Este trabalho tem por objetivo empregar algoritmos de predição junto com técnicas de pré-processamento de texto e analisar qual combinação produz melhores resultados no processo para prever o fator de impacto de notícias publicadas e também apresentar os motivos que podem exercer influência no fator de impacto sobre notícias divulgadas em páginas de fãs da rede social facebook.

As seções a seguir estão organizadas de forma a facilitar compreensão deste documento.

Capítulo 1: Apresenta uma introdução sobre a dissertação, o que motiva esta pesquisa e também uma sumarização dos objetivos da pesquisa.

Capítulo 2: Apresenta um resumo de Trabalhos Relacionados.

Capítulo 3: Apresenta a fundamentação teórica apropriada para a realização do trabalho.

Capítulo 4: Apresenta a metodologia evidenciando as tarefas utilizadas para atender ao objetivo do trabalho.

Capítulo 5: Apresenta uma discussão bem como uma análise sobre os resultados alcançados com a série de experimentos realizados utilizando os algoritmos mencionados no trabalho.

Capítulo 6: Apresenta os resultados finais dos experimentos como também a conclusão e trabalhos futuros.

1.1 Motivação

A IA tem contribuído muito na constante busca de resolver problemas de forma eficiente, através de soluções algorítmicas, metodologias de programação como também através de técnicas cada vez mais avançadas que se assemelham a capacidade do ser humano de resolver problemas.

Além disso a mineração de dados também desperta um crescente interesse tanto no meio cientifico como comercial. A extração de informações úteis através de conteúdos armazenadas em base de dados mostra-se promissora para diferentes domínio de aplicação.

Uma vez que há um investimento em publicidade e criação de páginas para a divulgação de anúncios e marcas, além de produtos, serviços e entretenimento beneficiando empresas através das redes sociais, o presente trabalho faz uso de técnicas de IA para predizer o fator de impacto de notícias divulgadas em cinco páginas de fãs mais acessadas da Rede Social Facebook no Brasil, com a intenção de revelar fatores que exercem influência nas interações, bem como os motivos que podem modificar o fator de impacto, através do processo de descoberta de conhecimento em base de dados e também fazendo uso de técnicas de processamento de linguagem natural com o objetivo de atender a expectativa desta dissertação além de identificar novas aplicações em comum dentro de Web Science.

1.2 Objetivos da Pesquisa

- ✓ Analisar e predizer o número de interações em notícias divulgadas em Redes Sociais.
- ✓ Determinar o Fator de Impacto de uma notícia divulgada em Redes Sociais.

- ✓ Identificar através de experimentos empregando Técnicas de Classificação. Pois podem ser utilizadas tanto para entender os dados existentes quanto para predizer o Fator de Impacto de uma notícia publicada na Rede Social.
- ✓ Revelar quais fatores que exercem influência para que uma notícia divulgada tenha um impacto maior na rede social.

2 TRABALHOS RELACIONADOS

Os trabalhos apresentados estão organizados em ordem cronológica, sendo o primeiro artigo, (Wüthrich. 1998), considerado o precursor na abordagem da predição do mercado financeiro através da análise de informações textuais. Também é apresentado um resumo de abordagens, o qual está organizado em seis subseções, onde se introduz o objetivo e técnicas empregadas nas respectivas tarefas e por fim o comparativo entre os trabalhos relacionados e a proposta deste trabalho.

2.1 Wüthrich et al. (1998)

O modelo proposto por (Wüthrich et al. 1998), tenta predizer a tendência diária de cinco índices da bolsa de valores, entre eles: o Dow Jones de Nova Iorque, o Nikkei de Tóquio, o FTSE de Londres, Hang Seng de Hong Kong e o Straits Index de Singapura. A predição diária é baseada em artigos publicados nos portais de notícias como, por exemplo: The Wall Street Jornal ou Reuters, até o momento de sua divulgação às 7h45min da manhã, horário de Hong Kong.

A arquitetura do modelo de predição é formada por cinco repositórios de dados, são eles: o primeiro contendo as notícias do dia da predição, o segundo contendo os valores de fechamento das ações do dia anterior ao dia da predição, o terceiro contendo as notícias dos dias anteriores, o quarto contendo os últimos cem valores de fechamento das ações e o quinto contendo mais de quatrocentas palavras-chave que podem ser simples ou até mesmo compostas por cinco palavras.

Acoplado a esses cinco repositórios existe um agente para realizar o download de toda essa informação e um gerador de regras para elaborara-las e aplica-las à predição. O funcionamento do modelo de predição segue os seguintes passos:

1- O número de ocorrências das palavras-chave, nas notícias de cada dia, é contado. Essa contagem é o resultado da correspondência das palavras-chave do repositório de dados com as palavras contidas nas notícias. A técnica de correspondência usada não diferencia letras maiúsculas de minúsculas, utiliza algoritmos de stemmingn e não precisa corresponder exatamente com as palavras contidas nas notícias, que mesmo assim, são consideradas correspondentes. Por exemplo: a palavra-chave "stock drop" é correspondente à frase "stocks have really dropped";
2- As ocorrências das palavras-chave são transformadas em pesos, assim, para cada dia, cada palavra chave recebe um peso. A abordagem utilizada é baseada em três componentes: frequência de termos, a discriminação de documentos e a normalização; esses três componentes irão gerar um peso entre 0 e 1 para as palavras-chave;
3- Através dos pesos das palavras-chave e dos valores de fechamento das ações, regras probabilísticas são geradas baseadas no artigo (Wüthrich, 1998) do mesmo autor;
4- As regras geradas são aplicadas nas notícias do dia predizendo se um índice como o Down Jones ir subir, descer ou permanecer estável em relação ao valor de fechamento do dia anterior. Os autores consideram que o índice permanece estável com uma variação menor do que 0.5% positiva ou negativa, acima de +0.5% é considerada uma alta no índice e abaixo de -0.5% uma baixa no índice. A predição gerada e disponibilizada na Internet às 07h45min da manhã, horário de Hong Kong.

2.2 Lavrenko et al. (2000)

O modelo desenvolvido por (Lavrenko, 2000), consegue predizer o comportamento das ações do mercado financeiro identificando as notícias específicas que podem influencia-lo.

O modelo, utiliza como dados de entrada a informação textual, no caso do artigo são: as notícias e as séries temporais referentes aos preços das ações. Utilizando a análise gráfica, são extraída tendência das séries temporais referentes aos preços das ações, ao mesmo tempo em que técnicas de recuperação da informação são usadas para extrair documentos relevantes das notícias, esses documentos são correlacionados com as tendências de preço de acordo com o mesmo momento em que ambos ocorreram. Com o correlacionamento estabelecido, são gerados modelos de linguagens para cada tipo de tendência, essa correlação é usada para predizer a tendência futura de uma ação quando uma nova notícia é publicada.

O artigo identifica três tipos de tendência: positiva, negativa e sem tendência. Uma tendência é definida pelo artigo como sendo um intervalo de tempo onde, neste intervalo,

existe uma predominância do aumento, diminuição ou estabilidade do preço da ação. Já os modelos de linguagem proveem um framework para a classificação de textos, por exemplo: palavras como perda, queda, falência estão associadas a uma tendência de baixa nos preços das ações, enquanto palavras como aquisição, aliança estão associadas a uma tendência de alta nos preços das ações.

Os autores, através do correlacionamento estabelecido entre tipos de tendência e modelos de linguagem, definiram cinco categorias de notícias, são elas: “Surge” quando a notícia gerou uma alta maior que +0.75% no preço da ação, “Slight+” uma alta entre +0.5% e +0.75%, “No Recommendation” uma volatilidade entre +0.5% e -0.5%, “Slight-” quando gerou um baixa entre -0.5% e -0.75% e “Plunge” uma baixa maior que - 0.75% no preço da ação.

Um classificador foi treinado utilizando a abordagem de Näıve Bayes. Assim que uma notícia é publicada, o modelo gera uma recomendação de compra para determinada ação somente se a notícia é classificada nas categorias “Surge” ou “Slight+”; gera uma recomendação de venda somente se a notícia é classificada nas categorias “Slight-” ou “Plunge”. Caso a notícia seja classificada na categoria “No Recommendation”, o modelo não gera recomendações.

2.3 Mittermayer (2004)

Este modelo busca predizer as tendências dos preços das ações do mercado acionário após a publicação de comunicados da imprensa, como fonte, os comunicados à imprensa ao invés das notícias baseiam-se numa maior confiabilidade, uma vez que todo o comunicado à imprensa segue regras rígidas de governança corporativa e fiscalização governamental.

A constituição do modelo NewsCATS consiste basicamente em três componentes, são eles: o primeiro componente é responsável pelo pré-processamento do texto e serve de entrada para o segundo componente, que utiliza o motor de classificação SVM light classifier, classificando em três categorias distintas, dentre elas: “GOOD NEWS”, “BAD NEWS” ou “NO MOVERS”. Já o terceiro componente recebe toda a informação e como saída sinais, indicando a tendência das ações.

2.4 Schumaker et al. (2009)

O modelo desenvolvido por (Schumaker, 2009) mostra como predizer o valor que uma determinada ação alcançará vinte minutos após a publicação de uma notícia. Um exemplo seria uma ação que no momento custa $10,00 dólares, então uma notícia é publicada, o

modelo, com isso, gera uma predição relatando que a ação A vai alcançar o preço de $11,00 dólares nos próximos vinte minutos. Para isso os autores utilizaram uma abordagem de aprendizagem de máquina supervisionada.

Nesse modelo são aplicadas três técnicas de análise textual, dentre elas: bag of words, frases nominais e entidades nomeadas, responsáveis em identificar os termos importantes das notícias e armazená-las em uma base de dados. Este modelo possui também outra base de dados com o propósito de armazenar a quotação das ações por minuto, no modelo a notícia é pré-processada por três técnicas distintas de análise textual e uma abordagem SVR é utilizada para gerar a predição. Três métricas diferentes são aplicadas nos resultados gerados pelo modelo com o objetivo de avaliá-lo.

2.5 Buza (2012)

No trabalho (Buza, 2012), concentra-se na análise de documentos que aparecem em blogs. O componente mais interessante deste protótipo de software permite prever o número de feedbacks que um documento de blog irá receber. Nos experimentos de predição, são utilizados vários algoritmos. Como contribuição adicional, foi publicado o conjunto de dados, a fim de motivar a investigação nesta área de crescente interesse.

O componente analítico do trabalho permite prever o número de retornos que um documento deverá receber nas próximas 24 horas. Para previsão de feedbacks, o foco foram documentos que estão em blogs de sites húngaros. Dado alguns documentos de blog que estão no passado, para o qual já sabemos quando e quantos retornos já receberam, a tarefa é prever quantos retornos uma publicação recente receberá nas próximas H horas.

Como resultado é evidenciado que o estado da arte dos modelos de regressão executam bem, e que eles superam modelos ingênuos substancialmente. Por outro lado, os resultados mostram que há espaço para melhorias, para o desenvolvimento de novos modelos para blogs e que o problema de previsão de comentário parece ser uma tarefa não trivial: para as técnicas utilizadas, em métodos de conjuntos específicos, que só conseguiram uma melhoria marginal.

2.6 Schmitt et al. (2013)

Conforme (Schmitt, 2013), o trabalho leva em consideração o feedback (neste caso o número de likes) provenientes dos usuários de uma rede social para determinar as classes no processo de classificação. Seu objetivo concentra-se em avaliar como se comportam diferentes classificadores quando utilizadas técnicas de mineração de dados para a predição da

popularidade das postagens de uma página do Facebook. E visa avaliar o desempenho de três classificadores de Aprendizagem de Máquina supervisionados (Naive Bayes, Sequential Minimal Optmization e C4.5) com a tarefa de classificar a popularidade de postagens.

Foram executados dois conjuntos de testes para cada classificador variando-se o número de classes. As postagens foram divididas em duas classes (“boa” e “ruim”) e três classes (“boa”, “média” e “ruim”) conforme a popularidade das mesmas tendo em vista o número de *likes* que cada postagem obteve.

2.7 Shoen et al. (2013)

Segundo (Shoen, 2013) As mídias sociais fornecem uma quantidade impressionante de dados sobre os usuários e suas interações, oferecendo assim novas oportunidades de investigação para cientistas de dados, economistas e estatísticos, entre outros. Indiscutivelmente, uma das linhas mais interessantes do trabalho é a de prever eventos futuros e desenvolvimentos a partir de dados de mídia social. No entanto, o trabalho atual é fragmentado e falta de abordagens de avaliação amplamente aceitos. Além disso, as primeiras técnicas surgiram muito recentemente, pouco se sabe sobre seu potencial global, limitações e aplicabilidade geral a diferentes domínios. Portanto, compreender melhor o poder e as limitações da mídia social previsão é de extrema importância. Argumenta-se que os modelos estatísticos parece ser a abordagem, mais frutífera a aplicar para fazer previsões a partir de dados de mídia social.

2.8 Comparativo entre Trabalhos Relacionados

Os artigos descritos neste trabalho apresentam semelhanças em suas arquiteturas, fundamentalmente, associadas ao modelo adotado por todos. Alguns dos modelos possuem dados do mercado financeiro, documentos formados por informações textuais extraídas da Internet, utilizam pré-processamento textual nos documentos, um algoritmo de classificação e aplicam regras que correlacionam os dados do mercado financeiro com as saídas produzidas pelos modelos para, somente então, revelar a predição da tendência. Diferem na escolha dos dados do mercado financeiro e das fontes de informações textuais que usam.

Para um melhor entendimento, informação textual e qualquer notícia divulgada pela imprensa, comunicados das empresas, comentários em blogs ou através do Twitter, enfim,

qualquer informação textual que possa ser extraída e utilizada nos modelos dos trabalhos relacionados.

Tabela 1 – Comparação entre trabalhos relacionadas e o trabalho proposto

	Autores	*Informação Textual*	*Algoritmo de Classificação*	*KDD*	*PLN*	*Predição*
I.	**Wüthrich et al. (1998)**	**Notícias do The Wall Street de Londres ou Reuters**	• **Rule-based** • **KNN** • **Back-propagation**	**Pré-processamento**	**Frequência de termos**	**Índice da Bolsa de Valores**
II.	**Lavrenko et al. (2000)**	**Notícias**	• **Naive Bayes**	**Pré-processamento**	**Correlação de palavras**	**Tendências de Ações**
III.	**Mittermayer (2004)**	**Comunicados à Imprensa**	• **SVM ligth classifier**	**Pré-processamento**		**Tendência de preços de ações**
IV.	**Schumaker et al. (2009)**	**Notícias**	• **SVR – suport vector regression**	**Pré-processamento**	**Bag-of-words Frases nominais Entidades nomeadas**	**Valor de Ação**
V.	**Buza (2012)**	**Documentos de Blogs Húngaros**	• **Regression trees** • **Neural Network** • **RBF Networks** • **KNN** • **Linear Regression** • **Ensemble Models**	**Pré-processamento**	**Frequência de termos**	**Número de feedbacks de blogs**
VI.	**Schmitt et al. (2013)**	**Postagens de páginas do facebook**	• **NaiveBayes** • **SVM** • **C4.5**	**Pré-processamento**		**Número de postagens**
VII.	Shoen et al. (2013)					
VIII.	**Schünke et al. (2015)**	**Noticias de páginas de fãs do Facebook**	• **NaiveBayes J48** • **AdaBoostM1** • **DecisionTable** • **RandomTree** • **AttributeSelectionClasifier** • **LWL**	**Seleção dos Dados Pré-processamento TransformaçãMineração e interpretação**	**Unigram Bigram Trigram N-gram**	**Fator de Impacto**

Este trabalho propõem determinar o fator de impacto de notícias publicadas na rede social facebook, através da média das interações curtir, comentar e compartilhar, também predizer o fator de impacto de notícias utilizando 7 algoritmos de classificação bem como identificar o algoritmo que apresenta o melhor resultado e por fim como o diferencial, apresentar os motivos que podem exercer influência no fator de impacto através do processo de Descoberta de Conhecimentos em Base de Dados e também fazendo uso de técnicas de Processamento de Linguagem Natural.

3 FUNDAMENTAÇÃO TEÓRICA

Neste capitulo será tratado a fundamentação teórica afim de estabelecer uma sequência dos temas abordados no presente trabalho.

As seções a seguir estão organizadas da seguinte forma.

Na seção 3.1 apresenta o conceito de predição, classificação, regressão linear.

Na seção 3.2 apresenta o conceito de mineração de dados.

Na seção 3.3 apresenta o conceito de Descoberta de Conhecimento em base de dados.

Na seção 3.4 apresenta o conceito de Processamento de Linguagem Natural, bem como as técnicas N-Grams, uniram, bigram e trigram.

3.1 Predição

Segundo o dicionário online de português o significado de predição é definido como ato de afirmar com convicção aquilo que poderá acontecer num momento futuro. Segundo (REZENDE, 2003) A predição consiste na generalização de exemplos (classificação, regressão).

Segundo (Dietterich et Al., 1998), um algoritmo de AM preditivo é uma função que, dado um conjunto de exemplos rotulados, constrói um estimador. O rótulo ou etiqueta toma valores num domínio conhecido. Se esse domínio for um conjunto de valores nominais, tem-se um problema de classificação, também conhecido como aprendizado de conceitos e o estimador gerado é um classificador. Se o domínio for um conjunto infinito e ordenado de valores tem-se um problema de regressão, que induz um regressor. Um classificador (ou regressor), por sua vez também é uma função que, dado um exemplo não rotulado, atribui esse exemplo a uma das possíveis classes (ou a um valor real).

Conforme (TAN et Al., 2006) O objetivo principal das técnicas de aprendizado de máquina é descobrir automaticamente regras gerais em grandes conjuntos de dados, que permitam extrair informações implicitamente representadas.

3.1.1 Classificação

A classificação permite determinar a qual classe um objeto pertence, dados os valores de um conjunto de atributos do objeto. Neste trabalho é proposto a classe fator de impacto que consiste no cálculo da média das interações de uma página de fãs da Rede Social Facebook. Este fator de impacto foi determinado de acordo com as seguintes classes definidas, Fator de Impacto Alto para mensagens com uma média de interações curtir, comentar e compartilhar

maior em relação as demais, Fator de Impacto Médio com a média de interações medianas e Fator de Impacto Baixo barra notícias com poucas interações.

A escolha dos algoritmos empregados neste trabalho se deve a tarefa de classificação desejada através da ferramenta Weka. Para comparar com outras estratégias de aprendizado de máquina, foram utilizadas técnicas baseadas em modelos de probabilidade condicional, modelo de árvore de decisão, modelo adaptativo, modelo de representações probabilísticas de alternativas, modelo de árvore aleatória formada por um processo estocástico, modelo que utiliza um classificador arbitrário e modelo baseado em instâncias.

3.1.2 Regressão Linear

Conforme (katti,2011), o objetivo do aprendizado supervisionado é induzir conceitos a partir de exemplos que estão pré-classificados, ou seja, exemplos que estão rotulados com uma classe conhecida. Se as classes possuem valores discretos, o problema é categorizado como classificação. Caso as classes possuam valores contínuos, o problema é categorizado como regressão. É supervisionado porque temos informações que conhecemos.

3.2 Mineração de Dados

Conforme (Fayyad et al.,1996), a mineração de dados é uma técnica de pesquisa que auxilia na busca de conhecimento que pertence a uma área mais ampla denominada descoberta de conhecimento, onde este processo envolve várias atividades que são divididas em três etapas principais identificadas como pré-processamento, mineração de dados e pós-processamento.

A etapa Data Mining (Mineração de Dados) é um processo onde são aplicados algoritmos para a extração de informações ou conhecimento implícito dos dados disponíveis.

Os dados armazenados contêm informações ocultas de grande relevância para o negócio. E devido ao grande volume de dados, a extração destas informações não é uma tarefa simples. E esta tarefa pertence a um processo conhecido como Descoberta de Conhecimento em Banco de Dados (DCBD) ou ainda na língua inglesa como Knowledge Discovery in Databases (KDD).

> Conforme (NAVEGA, 2002), É preciso ressaltar um detalhe que costuma passar despercebido na literatura: embora os algoritmos atuais sejam capazes de descobrir padrões "válidos e novos", ainda não temos uma solução eficaz para determinar padrões valiosos. Por essa razão, Data Mining ainda requer uma interação muito forte com analistas humanos, que são, em última instância, os principais responsáveis pela

> determinação do valor dos padrões encontrados. Além disso, a condução (direcionamento) da exploração de dados é também tarefa fundamentalmente confiada a analistas humanos, um aspecto que não pode ser desprezado em nenhum projeto que queira ser bem sucedido.

3.3 Descoberta de Conhecimento em Base de Dados

Para atender ao objetivo do trabalho são utilizados os passos do processo DCBD através da seleção dos dados, pré-processamento, formatação, mineração e interpretação com a finalidade de revelar algum conhecimento.

Segundo (katti,2011). A mineração de dados consiste em extrair ou "minerar" conhecimento a partir de grandes quantidades de dados. Em parte da literatura relacionada, a mineração de dados é também tratada como sinônimo para outro termo, a descoberta de conhecimento em base de dados (KDD, do inglês Knowledge Discovery in Databases). Outros autores consideram a mineração de dados uma etapa no processo de KDD, o qual compreende as seguintes etapas: a seleção, limpeza e integração dos dados, a transformação dos dados, a mineração dos dados e a avaliação e apresentação dos resultados (Fayyad et al.,1996).

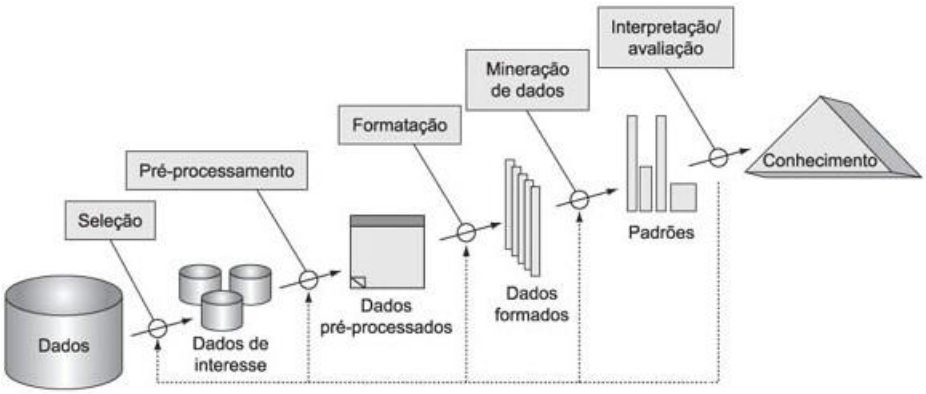


Figura 1 – Etapas do Processo KDD (Fayyad et al, 1996)

Conforme apresentado na Figura 1 acima o processo de Descoberta de Conhecimento descrito por (FAYYAD, 1996) consiste em selecionar os dados de interesse, realizar a etapa de pré-processamento dos dados, a formatação caso seja necessária, a etapa de mineração e por fim a interpretação que pode revelar conhecimento.

3.4 Processamento de Linguagem Natural

Conhecida como uma subárea da Inteligência Artificial o método de Processamento de Linguagem natural tem como objetivo principal estudar problemas de compreensão automática de línguas humanas, em outras palavras busca transformar informações armazenadas em bancos de dados em linguagem compreensível ao ser humano (JURAFSKY, 1999).

O NLTK é um conjunto de bibliotecas implementada na linguagem Python para o processamento de linguagem natural, com o objetivo de apoiar a pesquisa e o ensino em PNL e áreas afins incluindo a linguística, a ciência cognitiva, inteligência artificial, recuperação de informação como também aprendizagem de máquina. Incentivando a criação de protótipos bem como a construção de sistemas de apoio à pesquisa identificando características em textos ou ainda associação de textos com outros resultados.

E conforme (JURAFSKY, 1999) No processo de mineração de textos a ocorrência de palavras em sequência pode conter mais informação do que em palavras isoladas, permitindo desta forma, que se obtenha a probabilidade.

A probabilidade por sua vez permite que se calcule a chance de ocorrência de alguma coisa em um experimento aleatório afim de determinar a frequência. Antes de falar em probabilidade é preciso decidir o que vamos contar para calcular a probabilidade. A contagem em linguagem natural utiliza como base textos ou discursos também chamados de CORPUS. “Corpus é um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos).

3.4.1 N-Grams

N-GRAMS são junções ou combinações de palavras, onde N representa o número de palavras que foram unidas para a geração de um atributo. E essas palavras ou suas combinações são chamadas de n-grams. (JURAFSKY, 1999).

Com as combinações é possível gerar variações de atributos e essas derivações são denominadas unigram, bigram e trigram conforme apresentado na sequência.

3.4.2 Unigram

Unigram considera o termo como sendo um único elemento de texto, ou seja, uma única palavra. (JURAFSKY, 1999).

Tabela 2 – Exemplo da Técnica Unigram

Campo	*Unidade*	*Sequencia de Amostra*	*Unigram*
Sequencia Proteínas	Aminoácidos	...Cys-Gly-Leu-Ser-Trp...	...,Cys, Gly, Leu, Ser, Trp,...
Sequencia DNA	Pares de base	...AGCTTCGA...	..., A, G, C, T, T, C, G, A, ...
Linguística Computacional	Caracteres	...to_be_or_not_tobe...	..., t, o, _, b, e, _, o, r, _, n, o, t, _, t, o, b, e,...
Linguística Computacional	Palavras	..to be or not to be...	..., to, be, or, not, to, be, ...

3.4.3 Bigram

Bigrama considera o termo como sendo dois elementos de texto, ou seja, duas palavras. Normalmente esse termo é gerado pela concatenação de uma palavra com a próxima palavra do texto separada por um espaço em branco. (JURAFSKY, 1999).

Tabela 3 - Exemplo da Técnica Bigram

Campo	*Unidade*	*Sequencia de Amostra*	*Bigram*
Sequencia Proteínas	Aminoácidos	...Cys-Gly-Leu-Ser-Trp...	..., Cys-Gly, Gly-Leu, Leu-Ser, Ser-Trp, ...
Sequencia DNA	Pares de base	...AGCTTCGA...	..., AG, GC, CT, TT, TC, CG, GA, ...
Linguística Computacional	Caracteres	...to_be_or_not_tobe...	..., to, o_, _b, be, e_, _o, or, r_, _n, no, ot, t_, _t, to, o_, _b, be, ...
Linguística Computacional	Palavras	..to be or not to be...	..., to be, be or, or not, not to, to be, ...

3.4.4 Trigram

Trigrama considera o termo como sendo três elementos de texto, ou seja, três palavras. Isso significa que o termo é gerado pela concatenação de três palavras separadas por dois espaços em branco (JURAFSKY, 1999).

Tabela 4 - Exemplo da Técnica Trigram

Campo	*Unidade*	*Sequência de Amostra*	*Trigram*
Sequencia Proteínas	Aminoácidos	...Cys-Gly-Leu-Ser-Trp...	..., Cys-Gly-Leu, Gly-Leu-Ser, Leu-Ser-Trp, ...
Sequencia DNA	Pares de base	...AGCTTCGA...	..., AGC, GCT, CTT, TTC, TCG, CGA, ...
Linguística Computacional	Caracteres	...to_be_or_not_tobe...	..., to_, o_b, _be, be_, e_o, _or, or_, r_n, _no, not, ot_, t_t, _to, to_, o_b, _be, ...
Linguística Computacional	Palavras	..to be or not to be...	..., to be or, be or not, or not to, not to be, ...

A proposta de utilização das téquinicas de PLN apresentadas neste trabalho é de utiliza-las para verificar se há evidências de alguma influência, que justifique o aumento das interações, verificando a ocorrência de palavras ou ainda grupos de palavras que mais ocorrem nas notícias de divulgação de páginas de fãs da rede social facebook.

4 METODOLOGIA

Este capítulo apresenta as tecnologias empregadas, bem como, todas as atividades na aplicação de técnicas de PLN e DCBD, para revelar fatores que influenciam na predição de interações de redes sociais. Foram aplicadas as seguintes tecnologias:

- Facebook SDK for PHP;
- Linguagem de programação PHP (crawler);
- Facebook Query Language;
- Linguagem de banco de dados SQL;
- MySQL Server;
- Servidor Web Apache;
- Sistema Operacional Windows e Linux;
- Linguagem de Programação Phyton;
- Natural Language Toolkit Nltk;
- Weka.

4.1 Obtenção dos Dados da Página de Fãs do Facebook

Neste trabalho buscou-se informações que indicassem algumas páginas de fãs relevantes para o trabalho e segundo notícia divulgada no site Tecmundo é apresentada uma lista das Top 10 páginas fãs brasileiras mais curtidas no *facebook*, abaixo é apresentado o ranking com o número de pessoas que formam a base de fãs, observando que os números de fãs apresentado na lista abaixo é referente ao mês de dezembro de 2013.

- Guaraná Antarctica (15.938 milhões)
- Coca-Cola (15,750 milhões)
- Skol (12,935 milhões)
- Hotel Urbano (9,299 milhões)
- Garoto (9,104 milhões)
- Lacta (7,281 milhões)
- L`Oreal (6,430 milhões)
- Itaú (6,427 milhões)
- Halls (6,086 milhões)
- Netshoes (6,039 milhões)

Para o trabalho foi realizada a coleta, armazenamento e tratamento de informações referentes ao período de 2010 à 2014 extraídas de cinco página de fãs brasileiras presentes na lista citada acima, como Guaraná Antarctica, Coca-Cola, Hotel Urbano, Garoto e Lacta. Além disso para a realização da tarefa de obter os dados de uma página de fãs da rede social Facebook, foi utilizado o Facebook SDK que fornece um conjunto de funcionalidades que permite acessar chamadas de API do lado do servidor do Facebook incluindo todos os recursos da API Graph e FQL. Com o propósito de facilitar à extração de informações a utilizando para esta tarefa a tabela *stream* que possibilita a obtenção de 44 atributos de uma página de fãs.

4.2 Seleção dos Atributos da Página de Fãs do Facebook

A fim de, extrair informações uteis para a exploração de dados não estruturados foi realizado uma seleção, com o propósito de, identificar predicados pertinentes para responder ao objetivo do trabalho de prever o fator de impacto das interações de mensagens divulgadas

nas cinco páginas de fãs mais curtidas da rede social Facebook, e com isso, atender as expectativas do trabalho.

> Usando as próprias palavras do Facebook: As páginas de fãs ou fan pages existem para que as organizações, empresas, celebridades e bandas transmitam muitas informações aos seus seguidores ou ao público que escolher se conectar a elas. Semelhante aos perfis, as Páginas podem ser aprimoradas com aplicativos que ajudem as entidades a se comunicarem e interagirem com o seu público e adquirirem novos usuários por recomendações de amigos, históricos dos Feeds de notícias, eventos do Facebook e muito mais.

Conforme demonstrado na Figura 2 a seguir foram selecionados alguns atributos presentes na tabela *stream* visíveis nas páginas de fãs, entre eles é destacado o momento em que a notícia foi publicada, o conteúdo da mensagem além das interações número de curtidas, número de comentários e também número de compartilhamentos.

Figura 2 - Notícia divulgada na página de fãs da *Coca Cola*

Após a seleção dos atributos foram criados cinco bases de dados no sistema de gerenciamento de base de dados MYSQL, que permite utilizar a linguagem SQL para realizar operações de criação, leitura, atualização e exclusão dos dados com o objetivo de criar entidades para armazenar informações não estruturadas da web em atributos e possibilitar a exploração e extração de informações uteis.

Além de cada uma das bases de dados elaboradas, foram implementadas quatro novas entidades para cada página de fãs, com o propósito de armazenar as informações e prepará-las para a tarefa de pré-processamento. A primeira delas chamada de post contém 9 atributos selecionados da página de fãs do *Facebook*, já a entidade transformação_total contém 39 atributos obtidos na tarefa conhecida como pré-processamento, propondo atributos novos como por exemplo, o número de palavras presentes nas mensagens, como também no emprego de técnicas de processamento de linguagem natural aplicando as técnicas de unigram, bigram, trigram e ngram, que têm a finalidade de apresentar as cinco palavras com maior frequência presentes em cada corpus, bem como os cinco conjunto de duas palavras com maior frequência, como também os cinco conjuntos de três palavras com maior frequência, além de também o primeiro conjunto de cinco palavras com maior frequência e por fim primeiro conjunto de dez palavras com maior frequência, outros atributos como o mês, dia da semana, turno e o fator de impacto foram obtidos através dos atributos data e time minerados de postagens divulgadas das páginas de fãs. E ao final a entidade transformação_final contendo 56 atributos obtidos através da tarefa de classificação, que consiste em classificar os atributos de acordo com as seguintes classes sugeridas: alto, médio e baixo afim de verificar e evidenciar a existência de atributos que possam exercer influência na média das interações unificadas no atributo fator de impacto, conforme apresentado na tabela 5, e pôr fim a entidade predicao_final contendo os mesmos atributos da transformação_final, contudo nivelado o número de ocorrência de cada classes para precisar os resultados da predição.

Na sequência é apresentada a descrição de cada um dos atributos armazenados na entidade post, conforme apresentado na tabela 5.

Tabela 5 - Tabela POST Contendo Atributos Selecionados das páginas de fãs

Número	***Atributos/Predicados***	***Descrição***
1	post_id	O ID do post
2	message	A mensagem escrita no post
3	data	A data que o post foi publicado, expressa em UNIX timestamp
4	time	O tempo que o post foi publicado, expressa em UNIX timestamp
5	type	O tipo de história. Os valores possíveis são: 11 - Grupo criado 12 - Evento criado 46 - Atualização de status 56 - Mensagem na parede de outro usuário 66 - Nota criado 80 - Link postado 128 - Vídeo postado 247 - fotos postadas 237 - história App 257 - Comentário criado 272 - história App 285 - Entrada para um lugar 308 - Mensagem em Grupo
6	comment_info	Informações sobre os comentários deixados sobre este post (O número de comentários sobre este objeto.)
7	likes	Uma matriz contendo informações gostos (O número total de gostos)
8	share_count	Número de vezes que o post foi compartilhado
9	actor_id	ID do usuário que publicou o post

A partir dos atributos selecionados são sugeridos novos atributos conforme apresentado na tabela 6 com a finalidade de explorar conhecimento implícito presentes no conteúdo das notícias e na data que a notícia foi divulgada, para analisar se há indícios de influência dos novos atributos na classe Fator de Impacto.

Tabela 6 - Tabela POST Contendo Atributos Selecionados das páginas de fãs

Número	**Atributos**	**Detalhes dos novos atributos**
1	codigo	
2	post_id	
3	message	
4	palavras	**Contagem de palavras presentes em cada mensagem**
5	classe_palavras	
6	data	
7	ano	
8	mes	**Janeiro** **Fevereiro** **Março**

		Abril
		Maio
		Junho
		Julho
		Agosto
		Setembro
		Outubro
		Novembro
		Dezembro
9	dia_semana	**Domingo**
		Segunda-Feira
		Terça-Feira
		Quarta-Feira
		Quinta-Feira
		Sexta-Feira
		Sabado
10	time	
11	turno	**((Tempo >= 06:00:00) && (Tempo >= 11:59:59))** **Manhã**
		((Tempo >= 12:00:00) && (Tempo >= 17:59:59)) **Tarde**
		((Tempo >= 18:00:00) && (Tempo >= 23:59:59)) **Noite**
		((Tempo >= 00:00:00) && (Tempo >= 05:59:59)) **Madrugada**
12	comment_info	
13	classe_comentar	
14	likes	
15	classe_curtir	
16	share_count	
17	classe_compartilhar	
18	type	
19	type_descricao	11 Grupo criado
		12 Evento criado
		46 Atualização de status
		56 Mensagem na parede de outro usuário
		66 Nota criado
		80 Link postado
		128 Vídeo postado
		247 fotos postadas
		237 história App
		257 Comentário criado
		272 história App
		285 Entrada para um lugar
		308 Mensagem em Grupo
20	sentimento	
21	**umgram1**	A palavra mais frequente encontrada em cada corpus
22	classe_umgram1	
23	**umgram2**	A segunda palavra mais frequente encontrada em cada corpus
24	classe_umgram2	
25	**umgram3**	A terceira palavra mais frequente encontrada em cada corpus

26	classe_umgram3	
27	**umgram4**	A quarta palavra mais frequente encontrada em cada corpus
28	classe_umgram4	
29	**umgram5**	A quinta palavra mais frequente encontrada em cada corpus
30	classe_umgram5	
31	**bigram1**	O par de palavra mais frequente encontrada em cada corpus
32	classe_bigram1	
33	**bigram2**	O segundo par de palavra mais frequente encontrada em cada corpus
34	classe_bigram2	
35	**bigram3**	O terceiro par de palavra mais frequente encontrada em cada corpus
36	classe_bigram3	
37	**bigram4**	O quarto par de palavra mais frequente encontrada em cada corpus
38	classe_bigram4	
39	**bigram5**	O quinto par de palavra mais frequente encontrada em cada corpus
40	classe_bigram5	
41	**trigram1**	O conjunto de três palavra mais frequente encontrada em cada corpus
42	classe_trigram1	
43	**trigram2**	O segundo conjunto de três palavra mais frequente encontrada em cada corpus
44	classe_trigram2	
45	**trigram3**	O terceiro conjunto de três palavra mais frequente encontrada em cada corpus
46	classe_trigram3	
47	**trigram4**	O quarto conjunto de três palavra mais frequente encontrada em cada corpus
48	classe_trigram4	
49	**trigram5**	O quinto conjunto de três palavra mais frequente encontrada em cada corpus
50	classe_trigram5	
51	**ngram5**	O conjunto de cinco palavra mais frequente encontrada em cada corpus
52	classe_ngram5	
53	**ngram10**	O conjunto de dez palavra mais frequente encontrada em cada corpus
54	classe_ngram10	
55	**fator_impacto**	((nº curtidas + nº comentarios + nº compartilhamentos) /3)
56	**classe_fator_impacto**	**(Fator_impacto >= 66.67%)** — **Fator de impacto Alto** **((Fator_impacto >= 33.33%) && (Fator_impacto < 66.67%))** — Fator de impacto Médio **(Fator_impacto < 33.33%)** — Fator de impacto Baixo

4.3 Pré-Processamento dos Dados

Conforme (Katti, 2011) Técnicas de pré-processamento são frequentemente utilizadas para tornar os conjuntos de dados mais adequados para o uso de algoritmos AM. Essas técnicas podem ser agrupadas nos seguintes grupos de tarefas:

- Eliminação manual de atributos;
- Integração de dados;
- Amostragem de dados;
- Balanceamento de dados;
- Limpeza de dados;
- Redução de dimensionalidade;
- Transformação de dados;

E também segundo (Katti, 2011) As Técnicas de pré-processamento de dados são frequente utilizadas para melhorar a qualidade dos dados, facilitando o uso de técnicas de AM e possibilitar a construção de modelos mais fiéis à distribuição real dos dados.

Com a finalidade de evidenciar a etapa de pré-processamento, são apresentadas algumas atividades realizadas no corpus extraído das páginas de fãs, demonstrando as etapas realizadas no pré-processamento. Contudo devido à ambição do trabalho em realizar estas etapas em dados obtidos das cinco páginas de fãs mais curtidas no Brasil em 2013, como também apresentar uma análise dos resultados ao final do trabalho, é evidenciado que o problema concentra-se no volume de trabalho e tempo necessário para a realização desta fase. Considerada na literatura como exaustiva, tomando em média até 80% de todo tempo necessário para o processo completo. Na sequência são apresentados os seguintes passos realizados: definição dos objetivos, coleta de dados, pré-processamento dos dados e transformação de dados.

4.3.1 Remoção de Stop Words

Stop Words são conhecidas na literatura como palavras de parada, ou seja, segundo (Stanley, 2008) são palavras muito frequentes e com pouco significado como, por exemplo, exemplo artigos, preposições, algumas conjunções, normalmente são desconsideradas nas

minerações de texto. Para a realização desta tarefa foram implementados duas funções contendo uma lista de palavras de parada na língua Portuguesa e também na língua Inglesa a fim de removê-las no corpus, gerando um arquivo de saída sem Stop Words.

4.3.2 Remoção de Pontuação

As ocorrências de pontuações acabam transformando as palavras com a mesma sintaxe, ou seja a mesma forma de escrita. Como exemplo é citada a palavra concluído que acaba sendo considerada diferente da palavra concluído! Pois há ocorrência de pontuação junto a palavra, e para evitar esse problema é apresentada a implementação do método para realizar esta tarefa.

4.3.3 Remoção de Acentuação

O mesmo problema que temos nas pontuações, pode-se também, aplicar nas acentuações. Como exemplo é citada à palavra converção sendo diferente da palavra conversão. Desta forma, erros ortográficos no texto são facilmente confundidos com palavras diferentes, e para evitar esse problema é apresentado a implementação do método para realizar esta tarefa.

4.3.4 Palavras de Conteúdo

Após a realização da remoção das palavras de parada conhecidas como (Stop Words), remoção de pontuação e remoção de acentuação. O corpus estará preparado para aplicação das técnicas de processamento de linguagem natural como por exemplo unigram, bigram, trigram e ngram com o propósito de identificar palavras de conteúdo que ocorrem com maior frequência em cada um do corpus extraído de uma página de fãs. Logo palavras de conteúdo são conhecidas também como (content words), ou seja, são palavras pertencentes as categorias morfossintáticas como por exemplo: adjetivo, advérbio, numeral, interjeição, substantivo e algumas classificações somente incluem os verbos principais.

4.4 Criação do Corpora das Páginas de Fãs para calcular a Probabilidade

Para realizar os experimentos foram criados cinco corpus a partir da concatenação das mensagens presentes nas notícias extraídas de cada uma das cinco páginas de fãs da rede social *facebook*, gerando cinco arquivos.

Antes de falar em probabilidade é preciso decidir o que vamos contar para calcular a probabilidade. A contagem em linguagem natural utiliza como base textos ou discursos também chamados de CORPUS. “Corpus é um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos). A probabilidade permite que se calcule a chance de ocorrência de alguma coisa em um experimento aleatório afim de determinar a frequência. No processo de mineração de textos a ocorrências de palavras em sequência pode conter mais informação do que em palavras isoladas. E desse modo é possível criar atributos pela união de duas ou mais palavras consecutivas gerando atributos com um poder de predição maior. Como resultado cinco arquivos contendo dados minerados das cinco páginas de fãs mais curtidas da rede social Facebook do Brasil foram geradas, fornecendo a matéria prima para o trabalho.

4.4.1 Experimentos realizados utilizando a biblioteca NLTK

Uma vez constituídos os Corpora, como arquivos formados por conjuntos de mensagens através da concatenação de notícias extraídas de cada uma das páginas de fãs formando uma sequência longa de caracteres foi considerado explorar, toquens de cada corpus que constituem o texto. E de forma a ampliar os experimentos apresentando dez Unigram, dez Bigram, dez trigram, cinco e dez N-gram mais frequentes.

Com este experimento, objetiva-se demonstrar quais são as Unigram, Bigram, Trigram e N-Gram mais frequentes que ocorrem em cada um dos corpora utilizados neste trabalho, para posterior análise sobre a influência do Fator de Impacto.

4.4.1.1 Experimento 1 – Aplicação da Técnica Unigram

O proposito destes experimentos é revelar as dez palavras que mais ocorrem em todas as notícias de cada corpus mencionados no trabalho a fim de verificar se as palavras que mais ocorrem geram alguma influência no aumento de interações.

A fim de evidenciar os resultados dos experimentos são apresentadas as Unigram mais frequentes que aparecem nas notícias extraídas da página de fãs do *Guaraná Antarctica,* revelando quais são as dez Unigram mais frequentes, indicando quantas vezes aparece cada palavra nas postagens conforme visualizado na Figura 3.

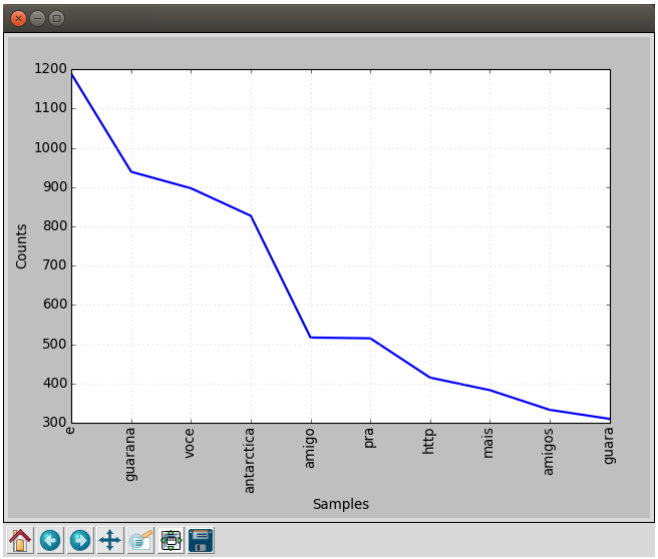


Figura 3 – Resultado da Técnica Unigram aplicado no corpus *Guaraná Antarctica*

Na sequência as Unigram mais frequentes que aparecem nas notícias extraídas da página de fãs da *Coca Cola*, indicando quantas vezes aparece cada palavra nas postagens conforme visualizado na Figura 4.

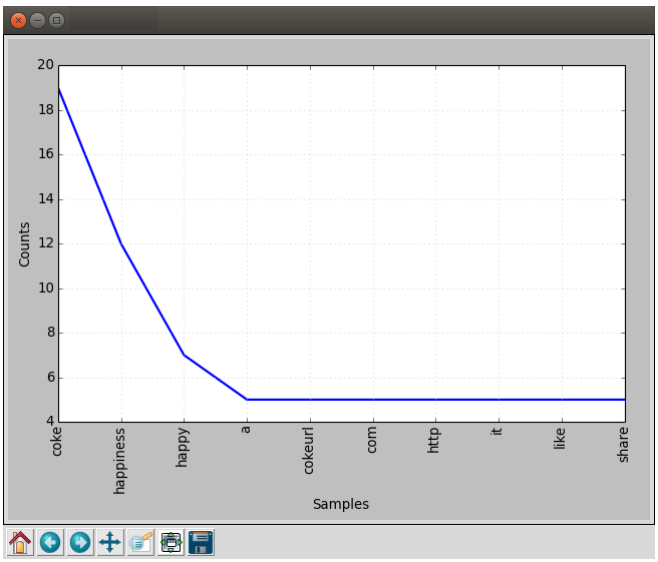


Figura 4 - Resultado da Técnica Unigram aplicado no corpus *Coca Cola*

Na ordem as Unigram mais frequentes que aparecem nas notícias extraídas da página de fãs do *Hotel Urbano*, são reveladas quais são as dez Unigram mais frequentes, indicando quantas vezes aparece cada palavra nas postagens conforme visualizado na Figura 5.

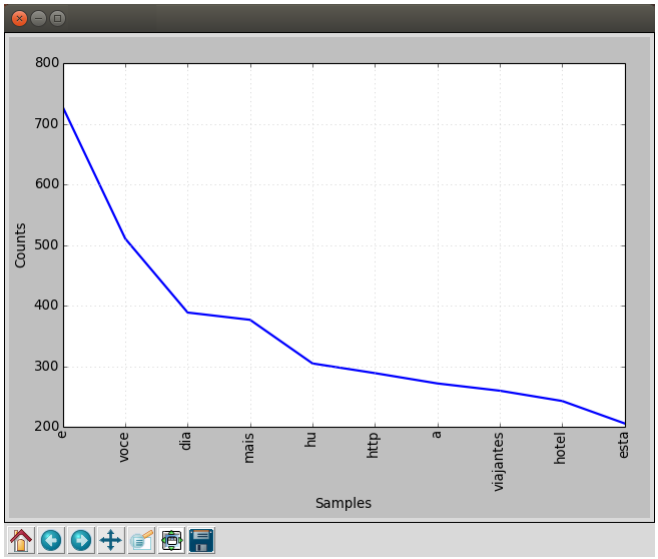


Figura 5 - Resultado da Técnica Unigram aplicado no corpus *Hotel Urbano*

No seguimento as Unigram mais frequentes que aparecem nas notícias extraídas da página de fãs da *Garoto*, é revelado quais são as dez Unigram mais frequentes, indicando quantas vezes aparece cada palavra nas postagens conforme visualizado na Figura 6.

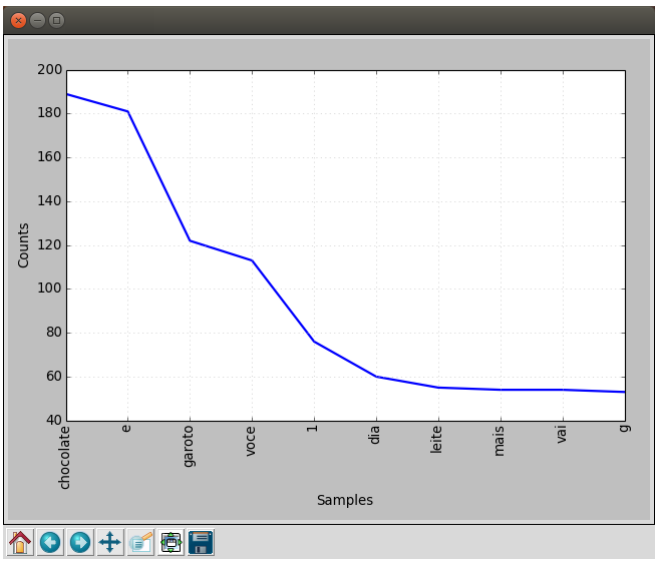


Figura 6 - Resultado da Técnica Unigram aplicado no corpus *Garoto*

Por conseguinte as Unigram mais frequentes que aparecem nas notícias extraídas da página de fãs da *Lacta*, é revelado quais são as dez Unigram mais frequentes, indicando quantas vezes aparece cada palavra nas postagens conforme visualizado na Figura 7.

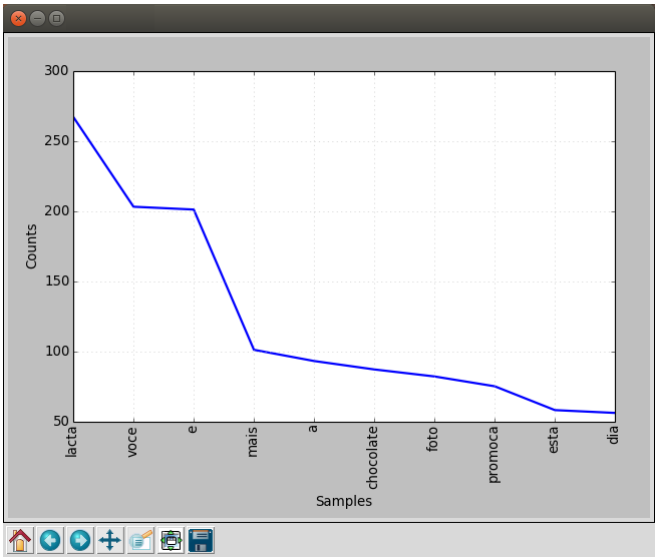


Figura 7 - Resultado da Técnica Unigram aplicado no corpus *Lacta*

Por fim as Unigram mais frequentes que aparecem nas notícias extraídas das páginas de fãs unificadas chamada de *União dos Copora*, é revelado quais são as dez Unigram mais frequentes, indicando quantas vezes aparece cada palavra nas postagens conforme visualizado na Figura 8.

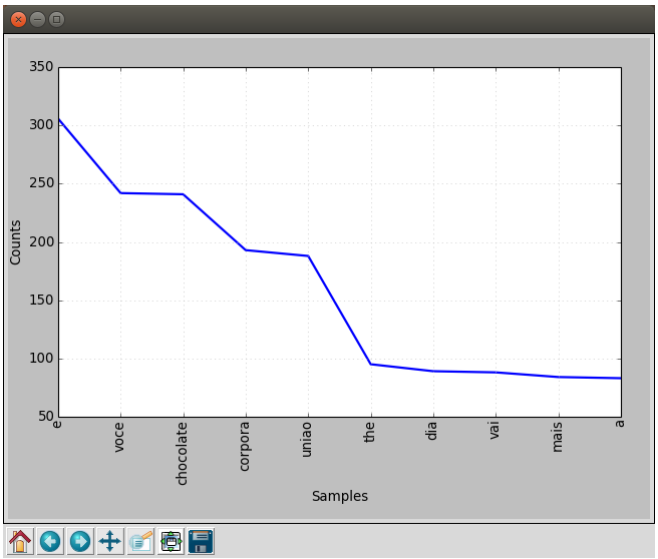


Figura 8 - Resultado da Técnica Unigram aplicado na *União dos Copora*

4.4.1.2 Experimento 2 – Aplicação da Técnica Bigram

O proposito destes experimentos é revelar os dez pares de palavras que mais ocorrem em todas as notícias de cada corpus mencionados no trabalho a fim de verificar se palavras que mais ocorrem em conjunto geram alguma influência no aumento de interações.

A fim de evidenciar os resultados dos experimentos é apresentado as Bigram mais frequentes que aparecem nas notícias extraídas da página de fãs do *Guaraná Antarctica*, é revelado quais são as dez Bigram mais frequentes, indicando quantas vezes aparece cada par de palavras nas postagens conforme visualizado na Figura 9.

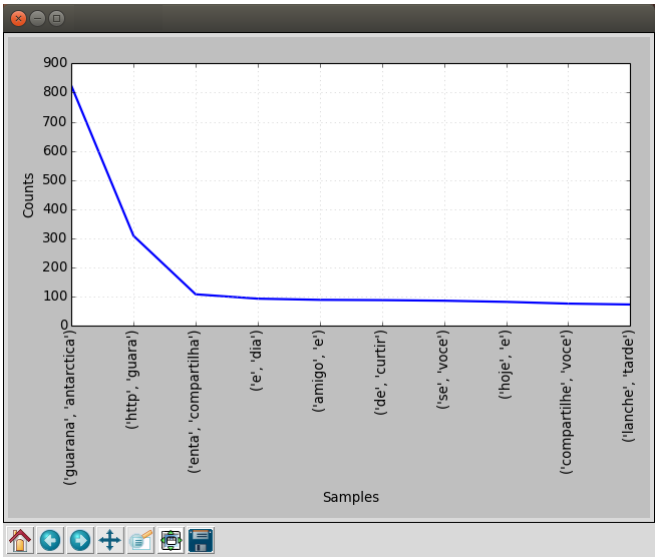


Figura 9 - Resultado da Técnica Bigram aplicado no corpus *Guaraná Antarctica*

Na sequência as Bigram mais frequentes que aparecem nas notícias extraídas da página de fãs da *Coca Cola*, é revelado quais são as dez Bigram mais frequentes, indicando quantas vezes aparece cada par de palavras nas postagens conforme visualizado na Figura 10.

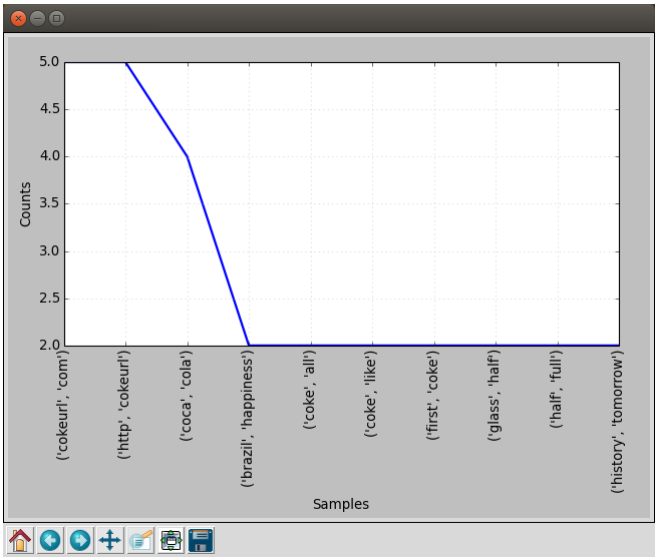


Figura 10 - Resultado da Técnica Bigram aplicado no corpus *Coca Cola*

Na ordem as Bigram mais frequentes que aparecem nas notícias extraídas da página de fãs do *Hotel Urbano*, é revelado quais são as dez Bigram mais frequentes, indicando quantas vezes aparece cada par de palavras nas postagens conforme visualizado na Figura 11.

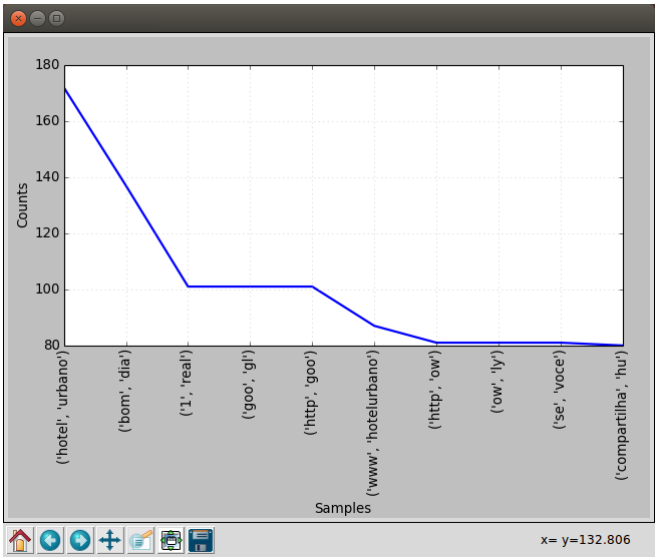


Figura 11 - Resultado da Técnica Bigram aplicado no corpus *Hotel Urbano*

No seguimento as Bigram mais frequentes que aparecem nas notícias extraídas da página de fãs da *Garoto*, é revelado quais são as dez Bigram mais frequentes, indicando quantas vezes aparece cada par de palavras nas postagens conforme visualizado na Figura 12.

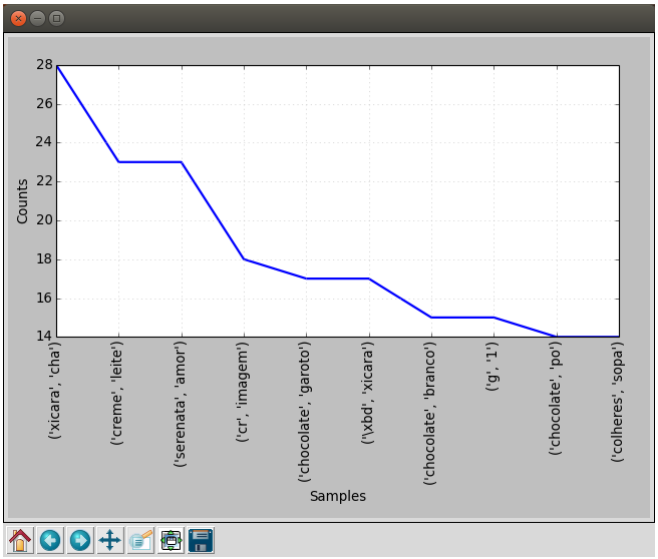


Figura 12 - Resultado da Técnica Bigram aplicado no corpus *Garoto*

Por conseguinte as Bigram mais frequentes que aparecem nas notícias extraídas da página de fãs da *Lacta*, é revelado quais são as dez Bigram mais frequentes, indicando quantas vezes aparece cada par de palavras nas postagens conforme visualizado na Figura 13.

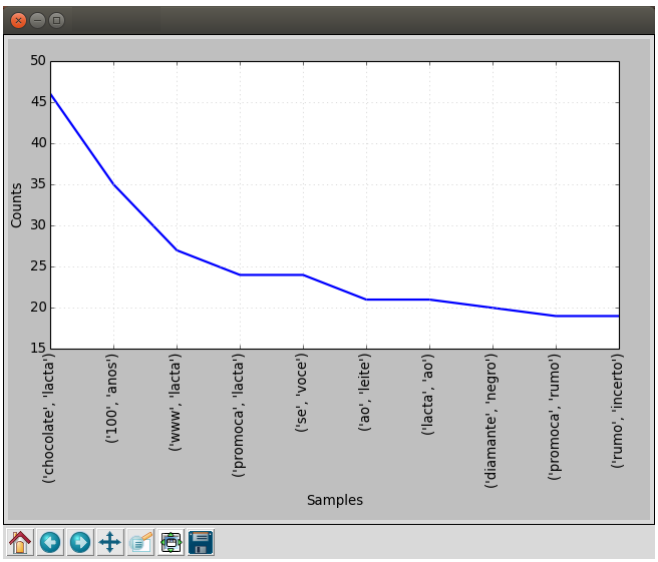


Figura 13 - Resultado da Técnica Bigram aplicado no corpus *Lacta*

Por fim as Bigram mais frequentes que aparecem nas notícias extraídas das páginas de fãs unificadas chamada de *União dos Copora*, é revelado quais são as dez Bigram mais frequentes, indicando quantas vezes aparece cada par de palavras nas postagens conforme visualizado na Figura 14.

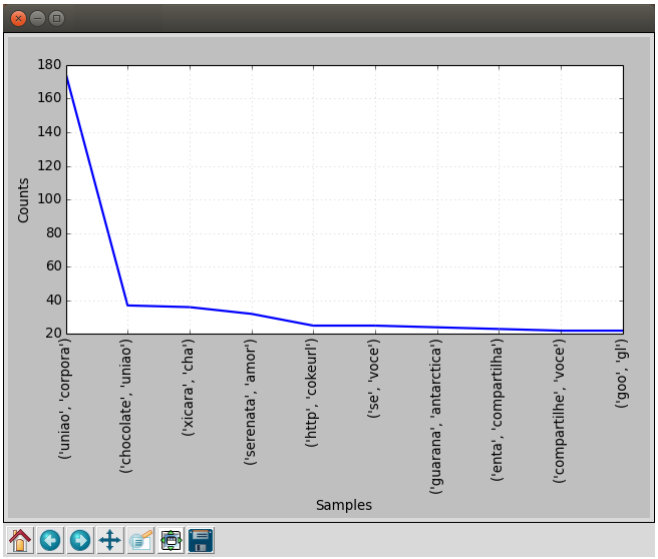


Figura 14 - Resultado da Técnica Bigram aplicado no corpus União dos *Copora*

4.4.1.3 Experimento 3 – Aplicação da Técnica Trigram

O proposito destes experimentos é revelar os dez conjuntos de três palavras que mais ocorrem em todas as notícias de cada corpus mencionados no trabalho a fim de verificar se estes conjuntos de palavras que mais ocorrem geram alguma influência no aumento de interações.

A fim de evidenciar os resultados dos experimentos é apresentado as Trigram mais frequentes que aparecem nas notícias extraídas da página de fãs do *Guaraná Antarctica*, é revelado quais são as dez Trigram mais frequentes, indicando quantas vezes aparecem cada conjunto de três palavras nas postagens conforme visualizado na Figura 15.

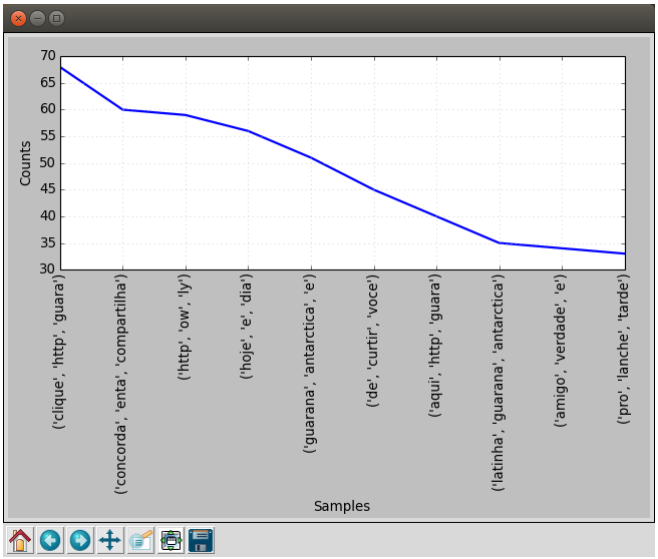


Figura 15 - Resultado da Técnica Trigram aplicado no corpus *Guaraná Antarctica*

Na sequência as Trigram mais frequentes que aparecem nas notícias extraídas da página de fãs da *Coca Cola*, é revelado quais são as dez Trigram mais frequentes, indicando quantas vezes aparecem cada conjunto de três palavras nas postagens conforme visualizado na Figura 16.

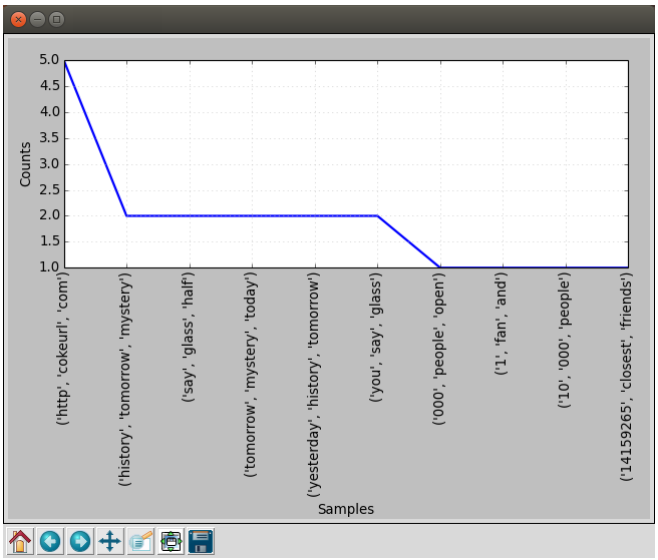


Figura 16 - Resultado da Técnica Trigram aplicado no corpus *Coca Cola*

Na ordem as Trigram mais frequentes que aparecem nas notícias extraídas da página de fãs da *Hotel Urbano*, é revelado quais são as dez Trigram mais frequentes, indicando quantas vezes aparecem cada conjunto de três palavras nas postagens conforme visualizado na Figura 17.

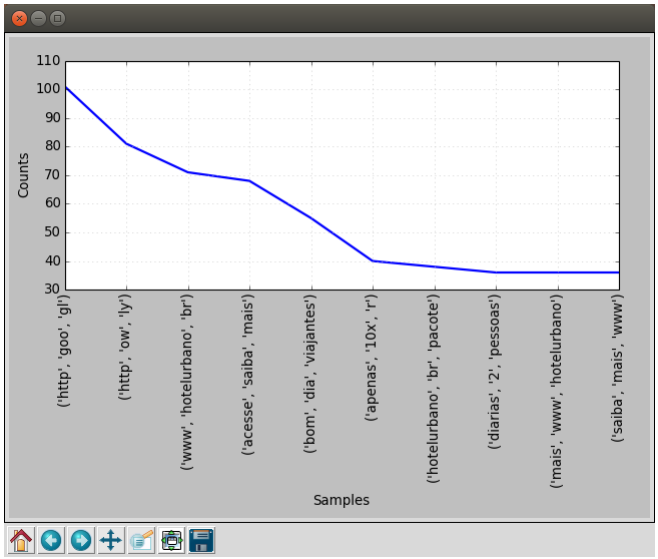


Figura 17 - Resultado da Técnica Trigram aplicado no corpus *Hotel Urbano*

No seguimento as Trigram mais frequentes que aparecem nas notícias extraídas da página de fãs da *Garoto*, é revelado quais são as dez Trigram mais frequentes, indicando quantas vezes aparecem cada conjunto de três palavras nas postagens conforme visualizado na Figura 18.

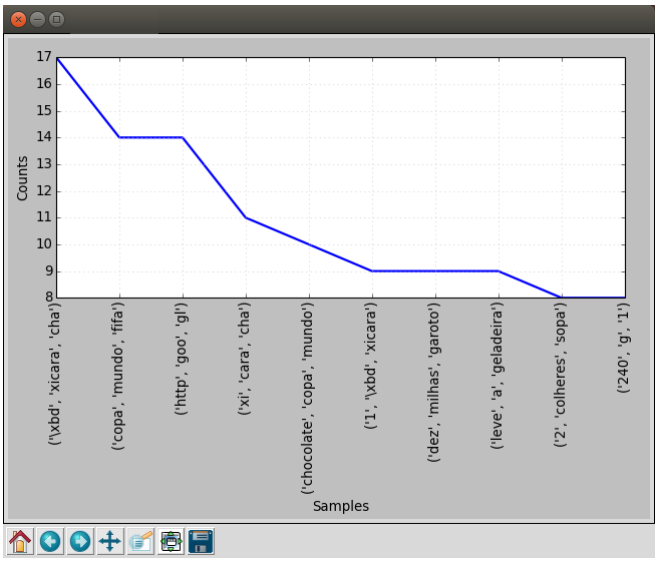


Figura 18 - Resultado da Técnica Trigram aplicado no corpus *Garoto*

Por conseguinte as Trigram mais frequentes que aparecem nas notícias extraídas da página de fãs da *Lacta*, é revelado quais são as dez Trigram mais frequentes, indicando quantas vezes aparecem cada conjunto de três palavras nas postagens conforme visualizado na Figura 19.

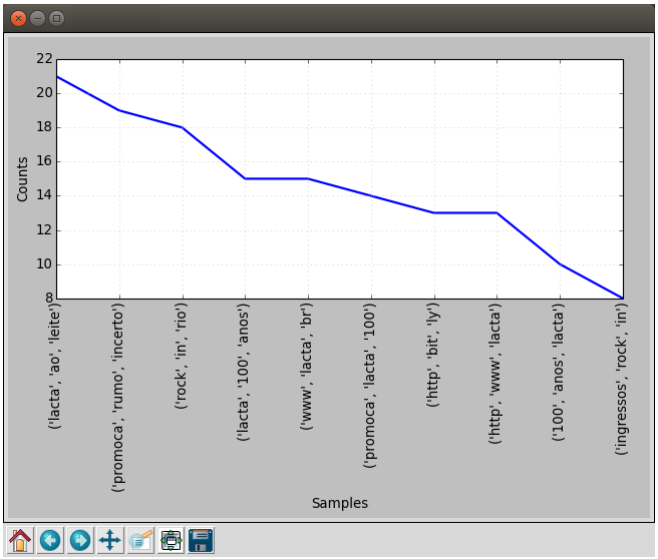


Figura 19 - Resultado da Técnica Trigram aplicado no corpus *Lacta*

Por fim as Trigram mais frequentes que aparecem nas notícias extraídas das páginas de fãs unificadas chamada de *União dos Copora*, é revelado quais são as dez Trigram mais frequentes, indicando quantas vezes aparecem cada conjunto de três palavras nas postagens conforme visualizado na Figura 20.

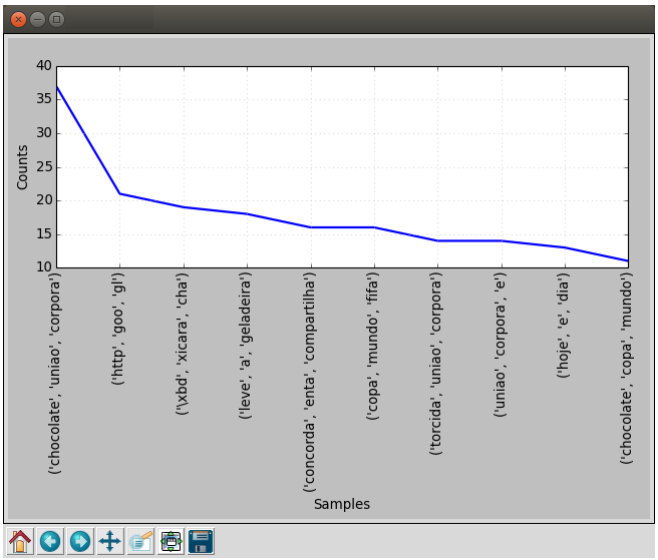


Figura 20 - Resultado da Técnica Trigram aplicado no corpus União dos *Copora*

4.5 Pós-Processamento dos Dados

Embora são apresentados como resultados as dez ocorrências das aplicações das técnicas Unigram, Bigram, Trigram e N-gram, definimos como ponto de corte utilizar as primeiras cinco ocorrências no presente trabalho com a intenção de utilizar os resultados mais relevantes observado nos gráficos apresentados.

Nesta etapa ainda não é possível uma analise sobre o conhecimento descoberto, uma vez que, o volume de informações obtidas é grande o que dificulta a análise. Para tanto, este trabalho prioriza também a revelação dos fatores que exercem influência sobre fator de impacto de uma notícia, sendo assim realizaremos a análise após a classificação das notícias por fator de impacto Alto, Médio e Baixo.

4.6 Classificação dos Atributos

Com a classificação é possível determinar a qual classe um objeto pertence, dados os valores de um conjunto de atributos do objeto. Por exemplo, neste trabalho é proposto o fator de impacto que consiste no cálculo da média das interações de uma página de fãs da Rede Social Facebook. Este fator de impacto foi definido e classificado de acordo com as seguintes classes, Fator de Impacto Alto, Fator de Impacto Médio e Fator de Impacto Baixo.

Abaixo é demonstrado um exemplo da classificação empregada no corpus obtido da página de fãs do Guaraná Antarctica conforme a fórmula.

$$Fi = \sum ((cu + co + cm) / 3)$$

Onde,

Fi Fator de impacto

cu Número de Curtidas de uma mensagem

co Número de Comentários de uma mensagem

cm Número de Compartilhamentos de uma mensagem

Exemplo:

cu = 999

co = 998

cm = 992

Fi = ∑(999 + 998 + 992) / 3

Fi = 996,33

Na tarefa de classificação é realizado uma consulta buscando o maior Fator de Impacto presente na base de dados, como exemplo foi identificado o maior fator de impacto encontrado na base das notícias mineradas da página de fãs do Guaraná Antarctica como:

Fi = 996,33

A partir do maior Fi encontrado é realizado a seguinte etapa onde,

Fa Fator de impacto alto

Fm Fator de impacto médio

Fb Fator de impacto baixo

Exemplo:

Fa = ((Fi * 66,67)/100)

Fa = ((996,33 * 66,67)/100)

Fa = 664,25

Fm = ((Fi * 33,33)/100)

Fm = ((996,33 * 33,33)/100)

Fm = 332,07

Uma vez calculado os limites das classes é realizada a classificação da mensagem

Se (Fi >= Fa)

Classe Fator de Impacto Alto

Se ((Fi >= Fm) && (Fi < Fa))

Classe Fator de Impacto Médio

Se (Fi < Fm)

Classe Fator de Impacto Baixo

Esta tarefa foi realizada da mesma forma sobre os dados da Coca-Cola, Hotel Urbano, Lacta, Garoto e também com a unificação de todas as notícias em uma única base chamada de União Corpora.

4.7 Análise da Predição

Neste trabalho foi definido a utilização do Weka, acrônimo de (Waikato Environment for Knowledge Analysis), uma ferramenta de mineração de dados que fornece uma coleção de algoritmos do estado da arte de aprendizagem de máquina além de ferramentas de pré-processamento, uma ferramenta de código aberto desenvolvida pala universidade de Waikato na Nova Zelândia.

Para a tarefa de predição são listados os algoritmos de classificação utilizados nesta área, bem como suas características.

- NaiveBayes (modelo de probabilidade condicional;)
- J48 (modelo de árvore de decisão)
- AdaBoostM1 (modelo adaptativo)
- DecisionTable (modelo de representações probabilisticas de alternativas)
- RandomTree (modelo de árvore aleatória formada por um processo estocástico)
- AttributeSelectedClassifier (modelo que utiliza um classificador arbitrário)
- LWL (modelo de aprendizado baseado em instâncias)

Conforme (BORGES, 2011) os principais benefícios do uso de diferentes algoritmos de classificação são para aumentar a qualidade do processo de redução de redundância, sem a necessidade humana intervenção e também compreender melhor as propriedades dos dados.

Outro fator da escolha da ferramenta é a facilidade de acesso e integração a diferentes fontes de dados como no caso deste trabalho acesso a base de dados Mysql, por meio de uma interface gráfica intuitiva.

Além disso foram definidos os seguintes parâmetros afim de melhorar os resultados:

IDFTransform = True
TFTransform = True
minTermFreq = 2
normalizeDocLength = Normalize all data
outputWordCount = True
stopwords = Weka-3-6

tokenizer = Alphabetic Tokenizer

useStoplist = True

words to keep = 1000

De modo que IDFTransform e TFTransform, quando outputWordCount = True, permitem a utilização da métrica de ponderação tf-idf nos dados, uma das medidas mais interessantes quando se lida com documentos de texto, pois permite dar maior relevância a palavras que não são comuns e que possam representar o documento. O parâmetro useStoplist, quando True, remove as stopwords dos candidatos a atributos (a lista de stopwords é dada no parâmetro stopwords). O Alphabetic Tokenizer utiliza apenas os tokens que são letras do alfabeto, removendo os outros símbolos (como @, #, +, -, ...). minTermFreq define a frequência mínima para manter uma palavra como candidata a atributo – neste caso, ela precisa ter frequência mínima 2. Finalmente, words to keep seleciona a quantidade máxima de palavras que cada classe pode manter, utilizando o default 1000.

O próximo filtro aplicado é o AttributeSelection, selecionando os atributos a serem utilizados pelo classificador. O parâmetro "evaluator" foi InfoGainAtributteEval, para buscar os atributos que melhor representam as classes.

Tabela 7 - Analise Estatística da Ocorrência do Fator de Impacto

Base	*Instâncias*	*Classe alto*	*Classe médio*	*Classe baixa*
Guaraná Antarctica	8344	4336	1223	2785
Coca-Cola	531	25	45	461
Hotel Urbano	7332	833	830	5669
Garoto	5023	185	441	4397
Lacta	2543	1005	187	1351
Uniao Corpora	23773	522	1390	21861

Foi necessário realizar o balanceamento dos dados, uma vez que não há um equilíbrio entre o número de ocorrências de cada classe.

Conforme observado na tabela 7, na base de dados do Guaraná Antarctica é identificada a ocorrência de 1223 mensagens classificadas como fator de impacto médio nos dados obtidos da página de fãs do Guaraná Antarctica e devido ao desbalanceamento das classes foi definido que o mesmo número de ocorrência encontrado nesta classe será utilizado a fim de equilibrar os dados para as classes alta e baixa com o objetivo de precisar os resultados da predição e tornar aceitável a acurácia preditiva do classificador para este conjunto de dados.

Afim de, evidenciar este processo é apresentado o histograma com as classes fator de impacto baixo, médio e alto precisamente balanceado conforme apresentado na Figura 21.

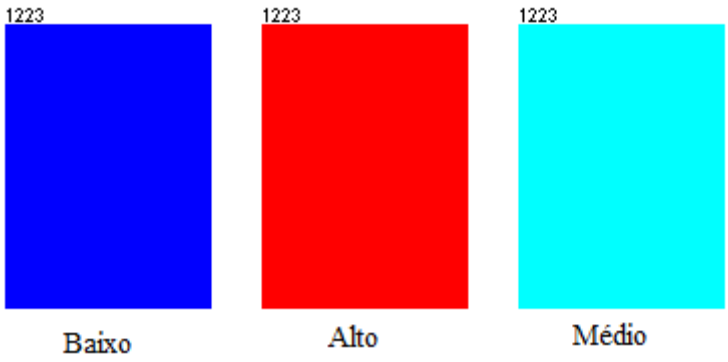


Figura 21 - Histograma congênere do fator de Impacto da base de dados *Guaraná Antarctica*.

Após o balanceamento entre as classes foi utilizado a ferramenta de mineração de dados Weka, bem como os seguintes algoritmos NaiveBayes, J48, AdaBoostM1, DecisionTable, RandomTree, AttributeSelectedClassifier e LWL.

Os resultados da predição utilizando cada um dos algoritmos de classificação citados, foram utilizados na tarefa de predizer o fator de impacto nos dados obtidos das cinco página de fãs mais acessadas no Brasil. Para esta tarefa foi utilizado o método de validação cruzada r-fold cross-validation, que utiliza uma conjunto de exemplos e os divide em r subconjuntos de tamanho aproximadamente iguais. O desempenho final do preditor é dado pela média dos desempenhos observados sobre cada subconjunto de teste.

Também foram considerados as seguintes métricas com o propósito de definir os critérios para avaliação afim de destacar a melhor técnica:

Precision: Instâncias classificadas corretamente como positivas dentre todos os que realmente são positivos.

Recall: Instâncias classificadas corretamente como positivos dentre todos que foram classificados como positivos.

F-Measure: Média ponderada entre Precisão e Recall, onde B é usado para ajustar a importância relativa entre Recall e Precision.

MCC: Coeficiente de Corelação Matthews leva em conta os verdadeiros e falsos positivos e negativos e é geralmente considerado como uma medida equilibrada, que pode ser usado mesmo se as classes são de tamanhos muito diferentes.

Basicamente o retorno possível na utilização destas métricas são valores entre -1 e + 1. Onde o coeficiente de + 1 representa uma previsão perfeita, 0 não é melhor que a previsão aleatória e -1 indica total desacordo entre previsão e observação.

Tabela 8 - Resultados da Classificação Utilizando o Algoritmo NaiveBayes

Resultados da Classificação Utilizando o Algoritmo NaiveBayes		
Correctly Classified Instances	3251	88.6072%
Incorrectly Classified Instances	418	11.3928%
Total Number of Instances	3699	

Tabela 9 – Precisão detalhada por classe utilizando NaiveBayes

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,942	0,942	0,942	0,851	
Médio	0,856	0,856	0,856	0,783	
Alto	0,866	0,866	0,866	0,857	
	0,888	0,886	0,866	0,830	Weighted Avg

O Algoritmo NaiveBayes utilizado neste experimento revelou um percentual de 88% de instâncias classificadas corretamente e 11% de instâncias classificadas de forma errada. Também é apresentado na Tabela 8 outras métricas, entre elas Acurácia, Precisão e Recall referem-se à assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 88%.

Tabela 10 - Resultados da Classificação Utilizando o Algoritmo J48

Resultados da Classificação Utilizando o Algoritmo J48		
Correctly Classified Instances	3373	91.9324%
Incorrectly Classified Instances	296	8.0676%
Total Number of Instances	3669	

Tabela 11 - Precisão detalhada por classe utilizando o Algoritmo J48

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,983	0,983	0,983	0,959	
Médio	0,951	0,951	0,951	0,834	
Alto	0,841	0,841	0,841	0,853	
	0,925	0,919	0,919	0,882	Weighted Avg

O Algoritmo J48 utilizado neste experimento revelou um percentual de 91% de instâncias classificadas corretamente e 8% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 10 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 91%.

Tabela 12 - Resultados da Classificação Utilizando o Algoritmo AdaBoostM1

Resultados da Classificação Utilizando o Algoritmo AdaBoostM1		
Correctly Classified Instances	2926	79.7493%
Incorrectly Classified Instances	743	20.2507%
Total Number of Instances	3699	

Tabela 13 - Precisão detalhada por classe utilizando o Algoritmo AdaBoostM1

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,966	0,966	0,966	0,940	
Médio	0,646	0,646	0,646	0,610	
Alto	0,862	0,862	0,862	0,584	
	0,825	0,797	0,793	0,711	Weighted Avg

O Algoritmo AdaBoostM1 utilizado neste experimento revelou um percentual de 79% de instancias classificadas corretamente e 20% de instâncias classificadas de forma errada. Também são apresentados na Tabela 12 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 79%.

Tabela 14 - Resultados da Classificação Utilizando o Algoritmo DecisionTable

Resultados da Classificação Utilizando o Algoritmo DecisionTable		
Correctly Classified Instances	3302	89.9973%
Incorrectly Classified Instances	367	10.0027%
Total Number of Instances	3669	

Tabela 15 - Precisão detalhada por classe utilizando o Algoritmo DecisionTable

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,958	0,958	0,958	0,936	
Médio	0,916	0,916	0,916	0,797	
Alto	0,836	0,836	0836	0,822	
	0,903	0,900	0,900	0,852	Weighted Avg

O Algoritmo DecisionTable utilizado neste experimento revelou um percentual de 89% de instancias classificadas corretamente e 10% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 14 outras métricas, entre elas Acurácia, Precisão e Recall referem-se à assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 89%.

Tabela 16 - Resultados da Classificação Utilizando o Algoritmo Random Tree

Resultados da Classificação Utilizando o Algoritmo RandomTree		
Correctly Classified Instances	1975	53.8294%
Incorrectly Classified Instances	1694	46.1706%
Total Number of Instances	3669	

Tabela 17 - Precisão detalhada por classe utilizando o Algoritmo RandomTree

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,487	0,487	0,487	0,366	
Médio	0,603	0,603	0,603	0,352	
Alto	0,600	0,600	0,600	0,255	
	0,563	0,538	0,520	0,324	Weighted Avg

O Algoritmo Random tree utilizado neste experimento revelou um percentual de 53% de instâncias classificadas corretamente e 46% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 16 outras métricas, entre elas Acurácia, Precisão e Recall referem-se à assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 53%.

Tabela 18 - Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier

Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier		
Correctly Classified Instances	3286	89.5612%
Incorrectly Classified Instances	383	10.4388
Total Number of Instances	3669	

Tabela 19 - Precisão detalhada por classe utilizando o Algoritmo AttributeSelectedClassifier

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,985	0,985	0,985	0,953	
Médio	0,992	0,992	0,992	0,803	
Alto	0,773	0,733	0,773	0,808	
	0,917	0,896	0,895	0,855	Weighted Avg

O Algoritmo AttributeSelectedClassifier utilizado neste experimento revelou um percentual de 89% de instancias classificadas corretamente e 10% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 18 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 89%.

Tabela 20 - Resultados da Classificação Utilizando o Algoritmo LWL

Resultados da Classificação Utilizando o Algoritmo LWL		
Correctly Classified Instances	3293	89.752%
Incorrectly Classified Instances	376	10.284%
Total Number of Instances	3669	

Tabela 21 - Precisão detalhada por classe utilizando o Algoritmo LWL

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,966	0,966	0,966	0,940	
Médio	0,948	0,948	0,948	0,790	
Alto	0,808	0,808	0,808	0,825	
	0,907	0,898	0,896	0,852	Weighted Avg

O Algoritmo LWL utilizado neste experimento revelou um percentual de 89% de instancias classificadas corretamente e 10% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 20 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 89%.

Também conforme observado na tabela 7 da pág 51, na base de dados da Coca-Cola é identificado a ocorrência de 25 mensagens classificadas como fator de impacto alto nos dados obtidos da página de fãs da Coca-Cola e devido a isso foi definido que o mesmo número de

ocorrência será utilizado afim de balancear os dados com o objetivo de precisar os resultados da predição.

Na sequência é apresentado o histograma com as classes fator de impacto baixo, médio e alto precisamente balanceado conforme apresentado na Figura 22.

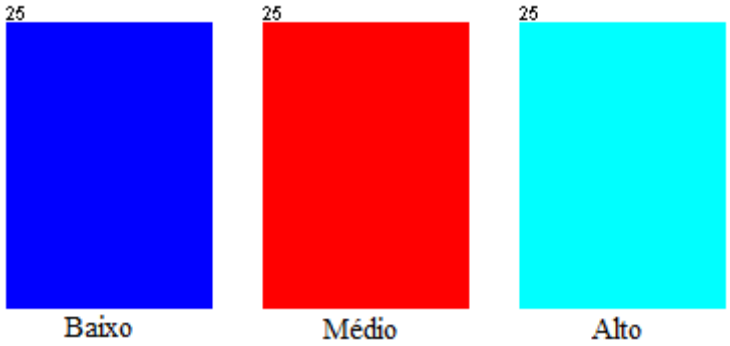


Figura 22 - Histograma congênere do fator de Impacto da base de dados *Coca Cola*

Na sequência os resultados da predição utilizando cada um dos algoritmos de classificação citados na base de dados extraídos da página de fãs da Coca-Cola.

Tabela 22 - Resultados da Classificação Utilizando o Algoritmo NaiveBayes

Resultados da Classificação Utilizando o Algoritmo NaiveBayes		
Correctly Classified Instances	42	56%
Incorrectly Classified Instances	33	44%
Total Number of Instances	75	

Tabela 23 – Resultado da Classificação utilizando NaiveBayes

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,810	0,810	0,810	0,630	
Médio	0,417	0,417	0,417	0,121	
Alto	0,500	0,500	0,500	0,289	
	0,575	0,560	0,564	0,347	Weighted Avg

O Algoritmo NaiveBayes utilizado neste experimento revelou um percentual de 56% de instancias classificadas corretamente e 44% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 22 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 56%.

Tabela 24 - Resultados da Classificação Utilizando o Algoritmo J48

Resultados da Classificação Utilizando o Algoritmo J48		
Correctly Classified Instances	40	53.3333%
Incorrectly Classified Instances	35	46.6667%
Total Number of Instances	75	

Tabela 25 - Precisão detalhada por classe utilizando o Algoritmo J48

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	1,000	1,000	1,000	0,882	
Médio	0,375	0,375	0,375	0,061	
Alto	0,333	0,333	0,333	0,000	
	0,569	0,533	0,548	0,314	Weighted Avg

O Algoritmo J48 utilizado neste experimento revelou um percentual de 53% de instancias classificadas corretamente e 46% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 24 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 53%.

Tabela 26 - Resultados da Classificação Utilizando o Algoritmo AdaBoostM1

Resultados da Classificação Utilizando o Algoritmo AdaBoostM1		
Correctly Classified Instances	41	54.6667%
Incorrectly Classified Instances	34	45.3333%
Total Number of Instances	75	

Tabela 27 - Precisão detalhada por classe utilizando o Algoritmo AdaBoostM1

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	1,000	1,000	1,000	0,882	
Médio	0,400	0,400	0,400	0,100	
Alto	0,345	0,345	0,345	0,019	
	0,582	0,547	0,561	0,334	Weighted Avg

O Algoritmo AdaBoostM1 utilizado neste experimento revelou um percentual de 54% de instancias classificadas corretamente e 45% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 26 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 54%.

Tabela 28 - Resultados da Classificação Utilizando o Algoritmo DecisionTable

Resultados da Classificação Utilizando o Algoritmo DecisionTable		
Correctly Classified Instances	54	72%
Incorrectly Classified Instances	21	28%
Total Number of Instances	75	

Tabela 29 - Precisão detalhada por classe utilizando o Algoritmo DecisionTable

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	1,000	1,000	1,000	0,882	
Médio	0,559	0,559	0,559	0,436	
Alto	0,700	0,700	0,700	0,469	
	0,753	0,720	0,726	0,596	Weighted Avg

O Algoritmo DecisionTable utilizado neste experimento revelou um percentual de 72% de instancias classificadas corretamente e 28% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 28 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 72%.

Tabela 30 - Resultados da Classificação Utilizando o Algoritmo RandomTree

Resultados da Classificação Utilizando o Algoritmo RandomTree		
Correctly Classified Instances	30	40%
Incorrectly Classified Instances	45	60%
Total Number of Instances	75	

Tabela 31 - Precisão detalhada por classe utilizando o Algoritmo RandomTree

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,409	0,409	0,409	0,191	
Médio	0,375	0,375	0,375	0,031	
Alto	0,391	0,391	0,391	0,082	
	0,392	0,400	0,360	0,101	Weighted Avg

O Algoritmo RandomTree utilizado neste experimento revelou um percentual de 40% de instancias classificadas corretamente e 60% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 30 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 40%.

Tabela 32 - Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier

Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier		
Correctly Classified Instances	41	54.6667%
Incorrectly Classified Instances	34	45.3333%
Total Number of Instances	75	

Tabela 33 - Precisão detalhada por classe utilizando o Algoritmo AttributeSelectedClassifier

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	1,000	1,000	1,000	0,882	
Médio	0,400	0,400	0,400	0,100	
Alto	0,345	0,345	0,345	0,019	
	0,582	0,547	0,561	0,334	Weighted Avg

O Algoritmo AttributeSelectedClassifier utilizado neste experimento revelou um percentual de 54% de instancias classificadas corretamente e 45% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 32 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 54%.

Tabela 34 - Resultados da Classificação Utilizando o Algoritmo LWL

Resultados da Classificação Utilizando o Algoritmo LWL		
Correctly Classified Instances	42	56%
Incorrectly Classified Instances	33	44%
Total Number of Instances	75	

Tabela 35 - Precisão detalhada por classe utilizando o Algoritmo LWL

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	1,000	1,000	1,000	0,882	
Médio	0,409	0,409	0,409	0,104	
Alto	0,375	0,375	0,375	0,076	
	0,595	0,560	0,572	0,354	Weighted Avg

O Algoritmo LWL utilizado neste experimento revelou um percentual de 56% de instancias classificadas corretamente e 44% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 34 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 56%.

Não só como também observado na tabela 6, na base de dados do Hotel Urbano é identificado a ocorrência de 830 mensagens classificadas como fator de impacto médio nos dados obtidos da página de fãs do Hotel Urbano e devido a isso foi definido que o mesmo número de ocorrência será utilizado afim de balancear os dados com o objetivo de precisar os resultados da predição.

Na sequência é apresentado o histograma com as classes fator de impacto baixo, médio e alto precisamente balanceado conforme apresentado na Figura 23.

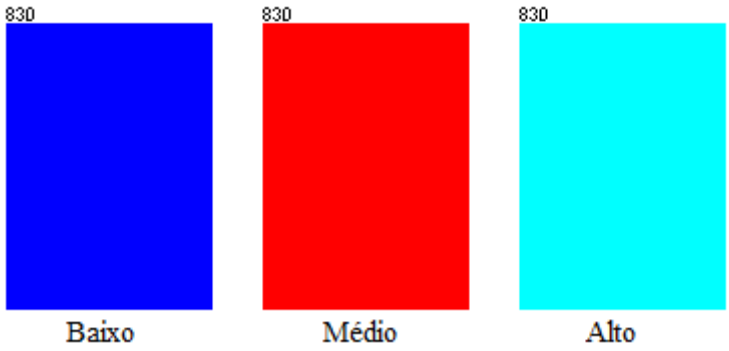


Figura 23 - Histograma congênere do fator de Impacto da base de dados *Hotel Urbano*

Na sequência os resultados da predição utilizando cada um dos algoritmos de classificação citados na base de dados extraídos da página de fãs do Hotel Urbano.

Tabela 36 - Resultados da Classificação Utilizando o Algoritmo NaiveBayes

Resultados da Classificação Utilizando o Algoritmo NaiveBayes		
Correctly Classified Instances	2172	87.2289%
Incorrectly Classified Instances	318	12.7711%
Total Number of Instances	2490	

Tabela 37 – Resultado da Classificação utilizando NaiveBayes

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,995	0,995	0,995	0,996	
Médio	0,784	0,784	0,784	0,722	
Alto	0,843	0,843	0,843	0,709	
	0,874	0,872	0,872	0,809	Weighted Avg

O Algoritmo NaiveBayes utilizado neste experimento revelou um percentual de 87% de instancias classificadas corretamente e 12% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 36 outras métricas, entre elas Acurácia, Precisão e Recall

referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 87%.

Tabela 38 - Resultados da Classificação Utilizando o Algoritmo J48

Resultados da Classificação Utilizando o Algoritmo J48		
Correctly Classified Instances	2105	84.5382%
Incorrectly Classified Instances	385	15.4618%
Total Number of Instances	2490	

Tabela 39 - Precisão detalhada por classe utilizando o Algoritmo J48

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	1,000	1,000	1,000	1,000	
Médio	0,757	0,757	0,757	0,656	
Alto	0,780	0,780	0,780	0,649	
	0,846	0,845	0,845	0,786	Weighted Avg

O Algoritmo J48 utilizado neste experimento revelou um percentual de 84% de instancias classificadas corretamente e 15% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 38 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 84%.

Tabela 40 - Resultados da Classificação Utilizando o Algoritmo AdaBoostM1

Resultados da Classificação Utilizando o Algoritmo AdaBoostM1		
Correctly Classified Instances	2086	83.7751%
Incorrectly Classified Instances	404	16.2248%
Total Number of Instances	2490	

Tabela 41 - Precisão detalhada por classe utilizando o Algoritmo AdaBoostM1

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,996	0,996	0,996	0,997	
Médio	0,679	0,679	0,679	0,702	
Alto	0,959	0,959	0,959	0,638	
	0,878	0,838	0,830	0,779	Weighted Avg

O Algoritmo AdaBoostM1 utilizado neste experimento revelou um percentual de 83% de instancias classificadas corretamente e 16% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 40 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 83%.

Tabela 42 - Resultados da Classificação Utilizando o Algoritmo DecisionTable

Resultados da Classificação Utilizando o Algoritmo DecisionTable		
Correctly Classified Instances	2119	85.1004%
Incorrectly Classified Instances	371	14.8996%
Total Number of Instances	2490	

Tabela 43 - Precisão detalhada por classe utilizando o Algoritmo DecisionTable

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,981	0,981	0,981	0,986	
Médio	0,779	0,779	0,779	0,668	
Alto	0,790	0,790	0,790	0,674	
	0,850	0,851	0,850	0,776	Weighted Avg

O Algoritmo DecisionTable utilizado neste experimento revelou um percentual de 85% de instancias classificadas corretamente e 14% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 42 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 85%.

Tabela 44 - Resultados da Classificação Utilizando o Algoritmo RandomTree

Resultados da Classificação Utilizando o Algoritmo RandomTree		
Correctly Classified Instances	1489	59.7992%
Incorrectly Classified Instances	1001	40.2008%
Total Number of Instances	2490	

Tabela 45 - Precisão detalhada por classe utilizando o Algoritmo RandomTree

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,540	0,540	0,540	0,502	
Médio	0,667	0,667	0,667	0,447	
Alto	0,688	0,688	0,688	0,313	
	0,632	0,598	0,572	0,421	Weighted Avg

O Algoritmo RandomTree utilizado neste experimento revelou um percentual de 59% de instancias classificadas corretamente e 40% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 44 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 59%.

Tabela 46 - Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier

Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier		
Correctly Classified Instances	2102	84.4177%
Incorrectly Classified Instances	388	15.5823%
Total Number of Instances	2490	

Tabela 47 - Precisão detalhada por classe utilizando o Algoritmo AttributeSelectedClassifier

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Alto	0,996	0,996	0,996	0,997	
Médio	0,756	0,756	0,759	0,653	
Baixo	0,780	0,780	0,780	0,649	
	0,844	0,844	0,844	0,766	Weighted Avg

O Algoritmo AttributeSelectedClassifier utilizado neste experimento revelou um percentual de 84% de instancias classificadas corretamente e 15% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 46 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 84%.

Tabela 48 - Resultados da Classificação Utilizando o Algoritmo LWL

Resultados da Classificação Utilizando o Algoritmo LWL		
Correctly Classified Instances	1886	75.743%
Incorrectly Classified Instances	604	24.257%
Total Number of Instances	2490	

Tabela 49 - Precisão detalhada por classe utilizando o Algoritmo LWL

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,996	0,996	0,996	0,997	
Médio	0,845	0,845	0,845	0,422	
Alto	0,586	0,586	0,586	0,574	
	0,809	0,757	0,733	0,664	Weighted Avg

O Algoritmo LWL utilizado neste experimento revelou um percentual de 75% de instancias classificadas corretamente e 24% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 48 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 75%.

Mas também conforme a tabela 7 da pág 51, na base de dados da Garoto é identificado a ocorrência de 185 mensagens classificadas como fator de impacto alto nos dados obtidos da página de fãs do Garoto e devido a isso foi definido que o mesmo número de ocorrência será utilizado afim de balancear a base de dados com o objetivo de precisar os resultados da predição.

Na sequência é apresentado o histograma com as classes fator de impacto baixo, médio e alto precisamente balanceado conforme apresentado na Figura 24.

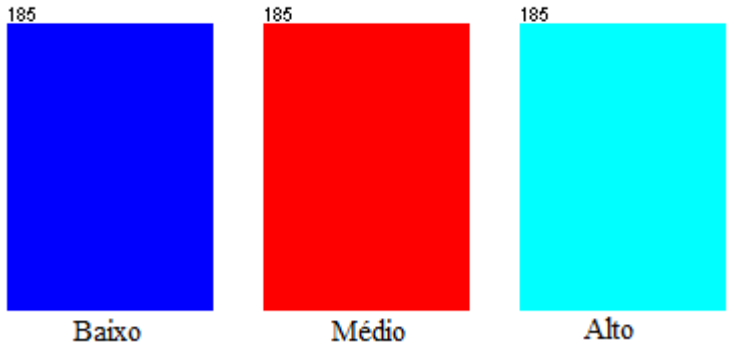


Figura 24 - Histograma congênere do fator de Impacto da base de dados *Garoto*

Na sequência os resultados da predição utilizando cada um dos algoritmos de classificação citados na base de dados extraídos da página de fãs do Garoto.

Tabela 50 - Resultados da Classificação Utilizando o Algoritmo NaiveBayes

Resultados da Classificação Utilizando o Algoritmo NaiveBayes		
Correctly Classified Instances	448	80.7207%
Incorrectly Classified Instances	107	19.2793%
Total Number of Instances	555	

Tabela 51 – Resultado da Classificação utilizando NaiveBayes

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,995	0,995	0,955	0,988	
Médio	0,699	0,699	0,699	0,573	
Alto	0,731	0,731	0,731	0,573	
	0,808	0,807	0,807	0,711	Weighted Avg

O Algoritmo NaiveBayes utilizado neste experimento revelou um percentual de 80% de instancias classificadas corretamente e 19% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 50 outras métricas, entre elas Acurácia, Precisão e Recall

referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 80%.

Tabela 52 - Resultados da Classificação Utilizando o Algoritmo J48

Resultados da Classificação Utilizando o Algoritmo J48		
Correctly Classified Instances	461	83.0631%
Incorrectly Classified Instances	94	16.9369%
Total Number of Instances	555	

Tabela 53 - Precisão detalhada por classe utilizando o Algoritmo J48

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,989	0,989	0,989	0,992	
Médio	0,665	0,665	0,665	0,698	
Alto	1,000	1,000	1,000	0,635	
	0,885	0,831	0,820	0,775	Weighted Avg

O Algoritmo J48 utilizado neste experimento revelou um percentual de 83% de instancias classificadas corretamente e 16% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 52 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 83%.

Tabela 54 - Resultados da Classificação Utilizando o Algoritmo AdaBoostM1

Resultados da Classificação Utilizando o Algoritmo AdaBoostM1		
Correctly Classified Instances	462	83.2432%
Incorrectly Classified Instances	93	16.7568%
Total Number of Instances	555	

Tabela 55 - Precisão detalhada por classe utilizando o Algoritmo AdaBoostM1

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	1,000	1,000	1,000	1,000	
Médio	0,668	0,668	0,668	0,701	
Alto	0,979	0,979	0,979	0,627	
	0,882	0,832	0,822	0,776	Weighted Avg

O Algoritmo AdaBoostM1 utilizado neste experimento revelou um percentual de 83% de instancias classificadas corretamente e 16% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 54 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 83%.

Tabela 56 - Resultados da Classificação Utilizando o Algoritmo DecisionTable

Resultados da Classificação Utilizando o Algoritmo DecisionTable		
Correctly Classified Instances	476	85.7658%
Incorrectly Classified Instances	79	14.2342%
Total Number of Instances	555	

Tabela 57 - Precisão detalhada por classe utilizando o Algoritmo DecisionTable

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,984	0,984	0,984	0,988	
Médio	0,710	0,710	0,710	0,736	
Alto	0,982	0,982	0,982	0,692	
	0,892	0,858	0,852	0,805	Weighted Avg

O Algoritmo DecisionTable utilizado neste experimento revelou um percentual de 85% de instancias classificadas corretamente e 14% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 56 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 85%.

Tabela 58 - Resultados da Classificação Utilizando o Algoritmo RandomTree

Resultados da Classificação Utilizando o Algoritmo RandomTree		
Correctly Classified Instances	301	54.2342%
Incorrectly Classified Instances	254	45.7658%
Total Number of Instances	555	

Tabela 59 - Precisão detalhada por classe utilizando o Algoritmo RandomTree

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,665	0,665	0,665	0,547	
Médio	0,449	0,449	0,449	0,194	
Alto	0,500	0,500	0,500	0,198	
	0,538	0,542	0,534	0,313	Weighted Avg

O Algoritmo RandomTree utilizado neste experimento revelou um percentual de 54% de instancias classificadas corretamente e 45% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 57 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 54%.

Tabela 60 - Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier

Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier		
Correctly Classified Instances	463	83.4234%
Incorrectly Classified Instances	92	16.5766%
Total Number of Instances	555	

Tabela 61 - Precisão detalhada por classe utilizando o Algoritmo AttributeSelectedClassifier

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	1,000	1,000	1,000	1,000	
Médio	0,668	0,668	0,668	0,708	
Alto	1,000	1,000	1,000	0,635	
	0,889	0,834	0,823	0,781	Weighted Avg

O Algoritmo AttributeSelectedClassifier utilizado neste experimento revelou um percentual de 83% de instancias classificadas corretamente e 16% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 60 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 83%.

Tabela 62 - Resultados da Classificação Utilizando o Algoritmo LWL

Resultados da Classificação Utilizando o Algoritmo LWL		
Correctly Classified Instances	435	78.3784%
Incorrectly Classified Instances	120	21.6216
Total Number of Instances	555	

Tabela 63 - Precisão detalhada por classe utilizando o Algoritmo LWL

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	1,000	1,000	1,000	1,000	
Médio	0,672	0,672	0,672	0,516	
Alto	0,680	0,680	0,680	0,511	
	0,784	0,784	0,784	0,676	Weighted Avg

O Algoritmo LWL utilizado neste experimento revelou um percentual de 78% de instancias classificadas corretamente e 21% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 62 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 78%.

Ainda conforme observado na tabela 7 pág. 51, na base de dados da Lacta é identificado a ocorrência de 187 mensagens classificadas como fator de impacto médio nos dados obtidos da página de fãs do Lacta e devido a isso foi definido que o mesmo número de ocorrência será utilizado afim de balanceado a base de dados para as classes baixa e alta com o objetivo de precisar os resultados da predição.

Na sequência é apresentado o histograma com as classes fator de impacto baixo, médio e alto precisamente balanceado conforme apresentado na Figura 25.

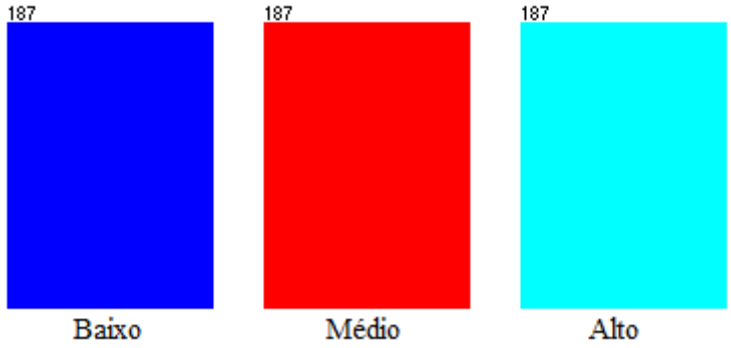


Figura 25 - Histograma congênere do fator de Impacto da base de dados *Lacta*

Na sequência os resultados da predição utilizando cada um dos algoritmos de classificação citados na base de dados extraídos da página de fãs da Lacta.

Tabela 64 - Resultados da Classificação Utilizando o Algoritmo NaiveBayes

Resultados da Classificação Utilizando o Algoritmo NaiveBayes		
Correctly Classified Instances	462	82.3529%
Incorrectly Classified Instances	99	17.6471%
Total Number of Instances	561	

Tabela 65 – Resultado da Classificação utilizando NaiveBayes

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,787	0,787	0,787	0,762	
Médio	0,779	0,779	0,779	0,652	
Alto	0,921	0,921	0,921	0,801	
	0,829	0,824	0,823	0,738	Weighted Avg

O Algoritmo NaiveBayes utilizado neste experimento revelou um percentual de 82% de instancias classificadas corretamente e 17% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 64 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 82%.

Tabela 66 - Resultados da Classificação Utilizando o Algoritmo J48

Resultados da Classificação Utilizando o Algoritmo J48		
Correctly Classified Instances	483	86.0963%
Incorrectly Classified Instances	78	13.9037%
Total Number of Instances	561	

Tabela 67 - Precisão detalhada por classe utilizando o Algoritmo J48

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,873	0,873	0,873	0,842	
Médio	0,831	0,831	0,831	0,703	
Alto	0,875	0,875	0,875	0,829	
	0,860	0,861	0,860	0,791	Weighted Avg

O Algoritmo J48 utilizado neste experimento revelou um percentual de 86% de instancias classificadas corretamente e 13% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 66 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 86%.

Tabela 68 - Resultados da Classificação Utilizando o Algoritmo AdaBoostM1

Resultados da Classificação Utilizando o Algoritmo AdaBoostM1		
Correctly Classified Instances	454	80.9269%
Incorrectly Classified Instances	107	19.0731%
Total Number of Instances	561	

Tabela 69 - Precisão detalhada por classe utilizando o Algoritmo AdaBoostM1

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,973	0,973	0,973	0,808	
Médio	0,656	0,656	0,656	0,635	
Alto	0,910	0,910	0,910	0,755	
	0,846	0,809	0,814	0,733	Weighted Avg

O Algoritmo AdaBoostM1 utilizado neste experimento revelou um percentual de 80% de instancias classificadas corretamente e 19% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 68 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 80%.

Tabela 70 - Resultados da Classificação Utilizando o Algoritmo DecisionTable

Resultados da Classificação Utilizando o Algoritmo DecisionTable		
Correctly Classified Instances	450	80.2139%
Incorrectly Classified Instances	111	19.7861%
Total Number of Instances	561	

Tabela 71 - Precisão detalhada por classe utilizando o Algoritmo DecisionTable

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,810	0,810	0,810	0,815	
Médio	0,765	0,765	0,765	0,603	
Alto	0,829	0,829	0,829	0,694	
	0,801	0,802	0,799	0,704	Weighted Avg

O Algoritmo DecisionTable utilizado neste experimento revelou um percentual de 80% de instancias classificadas corretamente e 19% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 70 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 80%.

Tabela 72 - Resultados da Classificação Utilizando o Algoritmo RandomTree

Resultados da Classificação Utilizando o Algoritmo RandomTree		
Correctly Classified Instances	246	43.8503%
Incorrectly Classified Instances	315	56.1497%
Total Number of Instances	561	

Tabela 73 - Precisão detalhada por classe utilizando o Algoritmo RandomTree

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,451	0,451	0,451	0,230	
Médio	0,382	0,382	0,382	0,078	
Alto	0,526	0,526	0,526	0,187	
	0,453	0,439	0,428	0,165	Weighted Avg

O Algoritmo RandomTree utilizado neste experimento revelou um percentual de 43% de instancias classificadas corretamente e 56% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 72 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 43%.

Tabela 74 - Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier

Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier		
Correctly Classified Instances	463	82.5312%
Incorrectly Classified Instances	98	17.4688%
Total Number of Instances	561	

Tabela 75 - Precisão detalhada por classe utilizando o Algoritmo AttributeSelectedClassifier

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,883	0,883	0,883	0,743	
Médio	0,722	0,722	0,722	0,642	
Alto	0,893	0,893	0,893	0,840	
	0,832	0,825	0,826	0,741	Weighted Avg

O Algoritmo AttributeSelectedClassifier utilizado neste experimento revelou um percentual de 82% de instancias classificadas corretamente e 17% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 74 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 82%.

Tabela 76 - Resultados da Classificação Utilizando o Algoritmo LWL

Resultados da Classificação Utilizando o Algoritmo LWL		
Correctly Classified Instances	459	82.3293%
Incorrectly Classified Instances	102	17.6707%
Total Number of Instances	561	

Tabela 77 - Precisão detalhada por classe utilizando o Algoritmo LWL

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,973	0,973	0,973	0,808	
Médio	0,696	0,696	0,696	0,641	
Alto	0,848	0,848	0,848	0,762	
	0,839	0,818	0,822	0,737	Weighted Avg

O Algoritmo LWL utilizado neste experimento revelou um percentual de 82% de instancias classificadas corretamente e 17% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 76 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 82%.

Por fim conforme observado na tabela 7 pág. 51, na base de dados *União dos Corpora* é identificado a ocorrência de 522 mensagens classificadas como fator de impacto alto sobre dados obtidos de todas as páginas de fãs e devido a isso foi definido que o mesmo número de ocorrência será utilizado afim de balanceado a base de dados para as classes baixa e alta com o objetivo de precisar os resultados da predição.

Na sequência é apresentado o histograma com as classes fator de impacto baixo, médio e alto precisamente balanceado conforme apresentado na Figura 26.

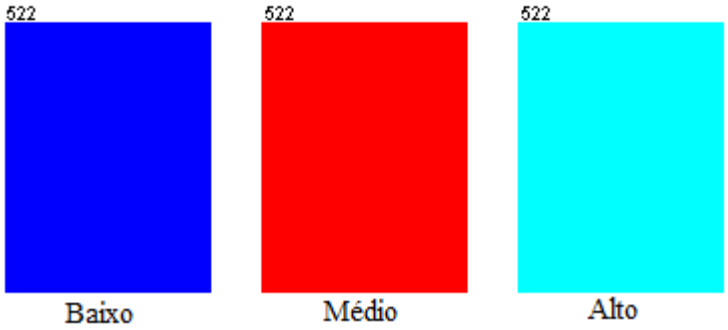


Figura 26 - Histograma congênere do fator de Impacto da base de dados da *União dos Corpora*

Na sequência os resultados da predição utilizando cada um dos algoritmos de classificação citados na base de dados extraídos de todas as páginas de fãs.

Tabela 78 - Resultados da Classificação Utilizando o Algoritmo NaiveBayes

Resultados da Classificação Utilizando o Algoritmo NaiveBayes		
Correctly Classified Instances	1233	75.8303%
Incorrectly Classified Instances	323	24.1697%
Total Number of Instances	1556	

Tabela 79 – Resultado da Classificação utilizando NaiveBayes

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,963	0,963	0,963	0,893	
Médio	0,640	0,640	0,640	0,535	
Alto	0,702	0,702	0,702	0,496	
	0,769	0,758	0,759	0,644	Weighted Avg

O Algoritmo NaiveBayes utilizado neste experimento revelou um percentual de 75% de instancias classificadas corretamente e 24% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 78 outras métricas, entre elas Acurácia, Precisão e Recall

referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 75%.

Tabela 80 - Resultados da Classificação Utilizando o Algoritmo J48

Resultados da Classificação Utilizando o Algoritmo J48		
Correctly Classified Instances	1103	67.8352 %
Incorrectly Classified Instances	453	32.1648 %
Total Number of Instances	1556	

Tabela 81 - Precisão detalhada por classe utilizando o Algoritmo J48

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,981	0,981	0,981	0,955	
Médio	0,524	0,524	0,524	0,306	
Alto	0,533	0,533	0,533	0,289	
	0,682	0,678	0,679	0,521	Weighted Avg

O Algoritmo J48 utilizado neste experimento revelou um percentual de 67% de instancias classificadas corretamente e 32% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 80 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 67%.

Tabela 82 - Resultados da Classificação Utilizando o Algoritmo AdaBoostM1

Resultados da Classificação Utilizando o Algoritmo AdaBoostM1		
Correctly Classified Instances	1111	68.3272%
Incorrectly Classified Instances	445	31.6728%
Total Number of Instances	1556	

Tabela 83 - Precisão detalhada por classe utilizando o Algoritmo AdaBoostM1

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,889	0,889	0,889	0,895	
Médio	0,659	0,659	0,659	0,240	
Alto	0,539	0,539	0,539	0,482	
	0,699	0,683	0,640	0,540	Weighted Avg

O Algoritmo AdaBoostM1 utilizado neste experimento revelou um percentual de 68% de instancias classificadas corretamente e 31% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 82 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 68%.

Tabela 84 - Resultados da Classificação Utilizando o Algoritmo DecisionTable

Resultados da Classificação Utilizando o Algoritmo DecisionTable		
Correctly Classified Instances	1271	78.1673%
Incorrectly Classified Instances	285	21.8327%
Total Number of Instances	1556	

Tabela 85 - Precisão detalhada por classe utilizando o Algoritmo DecisionTable

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,968	0,968	0,968	0,958	
Médio	0,733	0,733	0,733	0,496	
Alto	0,653	0,653	0,653	0,574	
	0,787	0,782	0,778	0,678	Weighted Avg

O Algoritmo DecisionTable utilizado neste experimento revelou um percentual de 78% de instancias classificadas corretamente e 21% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 84 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 78%.

Tabela 86 - Resultados da Classificação Utilizando o Algoritmo RandomTree

Resultados da Classificação Utilizando o Algoritmo RandomTree		
Correctly Classified Instances	771	47.417%
Incorrectly Classified Instances	785	52.583%
Total Number of Instances	1556	

Tabela 87 - Precisão detalhada por classe utilizando o Algoritmo RandomTree

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,455	0,455	0,455	0,387	
Médio	0,511	0,511	0,511	0,182	
Alto	0,548	0,548	0,548	0,146	
	0,504	0,474	0,411	0,240	Weighted Avg

O Algoritmo RandomTree utilizado neste experimento revelou um percentual de 47% de instancias classificadas corretamente e 52% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 86 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 47%.

Tabela 88 - Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier

Resultados da Classificação Utilizando o Algoritmo AttributeSelectedClassifier		
Correctly Classified Instances	1111	68.3272%
Incorrectly Classified Instances	445	31.6728%
Total Number of Instances	1556	

Tabela 89 - Precisão detalhada por classe utilizando o Algoritmo AttributeSelectedClassifier

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,991	0,991	0,991	0,962	
Médio	0,523	0,523	0,523	0,392	
Alto	0,563	0,563	0,563	0,231	
	0,695	0,683	0,666	0,534	Weighted Avg

O Algoritmo AttributeSelectedClassifier utilizado neste experimento revelou um percentual de 68% de instancias classificadas corretamente e 31% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 88 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 68%.

Tabela 90 - Resultados da Classificação Utilizando o Algoritmo LWL

Resultados da Classificação Utilizando o Algoritmo LWL		
Correctly Classified Instances	1074	66.0517%
Incorrectly Classified Instances	482	33.9483%
Total Number of Instances	1556	

Tabela 91 - Precisão detalhada por classe utilizando o Algoritmo LWL

Class	*Precision*	*Recall*	*F-Measure*	*MCC*	
Baixo	0,965	0,965	0,965	0,908	
Médio	0,506	0,506	0,506	0,470	
Alto	0,705	0,705	0,705	0,162	
	0,726	0,661	0,590	0,520	Weighted Avg

O Algoritmo LWL utilizado neste experimento revelou um percentual de 66% de instancias classificadas corretamente e 33% de instâncias classificadas de forma errada. Também são apresentadas na Tabela 90 outras métricas, entre elas Acurácia, Precisão e Recall referem-se a assertividade na previsão da classificação de instâncias corretas o que é evidenciado com 66%.

5 DISCUSSÃO E ANALISE DE RESULTADOS

Primeiramente foi realizada uma comparação entre os algoritmos de predição empregados no trabalho a fim de identificar qual algoritmo conseguiu maior desempenho na classificação de instâncias corretas. Numa primeira análise podemos observar que os melhores resultados da classificação do Fator de Impacto em todos os corpora utilizados no trabalho tiveram uma taxa de classificação de instâncias corretas considerados alta comparada ao estado da arte.

No entanto os melhores resultados da classificação ocorreram em algoritmos diferentes. Uma das hipóteses para justificar este comportamento é devido à variação do número de instâncias em cada corpus, devido ao fato da limitação do volume de dados disponível nas páginas de fãs no período de 2010 a 2014, mesmo se tratando de páginas de fãs mais curtidas no Brasil.

Em segunda análise pode-se observar que para este cenário especificamente, devido à variação do número de instâncias ocasionando a limitação do volume de dados de cada um dos corpora utilizados, é identificado que o algoritmo DecisionTable obteve um resultado equilibrado considerando também o segundo melhor resultado evidenciando um desempenho significativo em quatro corpora dos cincos utilizados. O mesmo também no corpora com as notícias de todas as páginas de fãs juntas conforme observado na tabela 92.

Tabela 92 – Resultados da Classificação do Fator de Impacto dos Algoritmos

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
NaiveBayes	88,6072%	56%	87,2289%	80,7207%	82,3529%	75,8303%
J48	91,9324%	53%	84,5382%	83,0631%	86,0963%	67,8352%
AdaBoostM1	79,7493%	54%	83,7751%	83,2432%	80,9269%	68,3272%
DecisionTable	89,9973%	72%	85,1004%	85,7658%	80,2139%	78,1673%
RandomTree	53,8294%	40%	59,7992%	54,2342%	43,8503%	47,417%
AttributeSelectedClassifier	89,5612%	54,6667%	84,4177%	83,4234%	82,5312%	68,3272%
LWL	89,752%	56%	75,743%	78,3784%	82,3293%	66,0517%
FATOR DE IMPACTO	3699	75	2490	555	561	1626

Também na tabela 92 é evidenciado que o algoritmo J48 obteve a melhor taxa de predição de 91,9324% no corpus do Guaraná Antarctica e também de 86,0963% considerada a terceira melhor no corpus da Lacta, embora um resultado significativo devido à ótima classificação do algoritmo nesse corpus observou que o mesmo resultado não foi evidenciado nos corpora da Coca Cola, Hotel Urbano, Garoto e muito menos na base de dados com mensagens de todas as páginas de fãs conforme também observado na Figura 27.

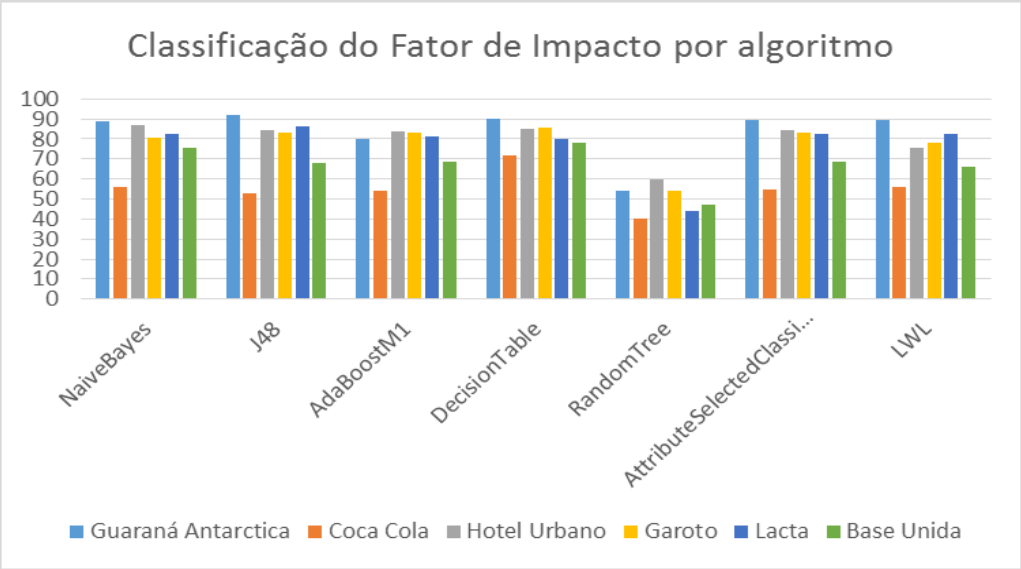


Figura 27 – Gráfico do Resultado da Classificação dos Corpora

Como tentativa de aumentar o volume de instâncias utilizadas para responder de forma satisfatória as questões do trabalho é experimentado a união dos dados minerados armazenados em cada uma das cinco bases de dados em uma única base de dados denominada uniao_corpora a fim de comparar o resultado apresentados na tabela 91 além de identificar o algoritmo que obteve a melhor taxa de classificação para determinar o fator de impacto de cada notícia divulgada.

Tabela 93 - Resultados da Classificação do Fator de Impacto com a Base Unificada

	Guaraná Antarctica, Coca Cola, Hotel Urbano, Garoto, Lacta
NaiveBayes	75,8303%
J48	67,8352%
AdaBoostM1	68,3272%
DecisionTable	78,1673%
RandomTree	47,417%
AttributeSelectedClassifier	68,3372%
LWL	66,0517%
INSTANCIAS	1626

Curiosamente no cenário onde foram unificadas as mensagens divulgadas das páginas de fãs mencionadas, conforme tabela 93, é evidenciado que o algoritmo que gerou melhor resultado na tarefa de classificação foi o DecisionTable, sugerindo que este algoritmo é adaptável tanto à variação do número de instâncias utilizadas na tarefa de classificação indicando que o volume de dados pode não influenciar no resultado. É constatado também que 78,1673% é uma taxa de predição. Na sequência a Figura 28 apresenta outra perspectiva dos resultados.

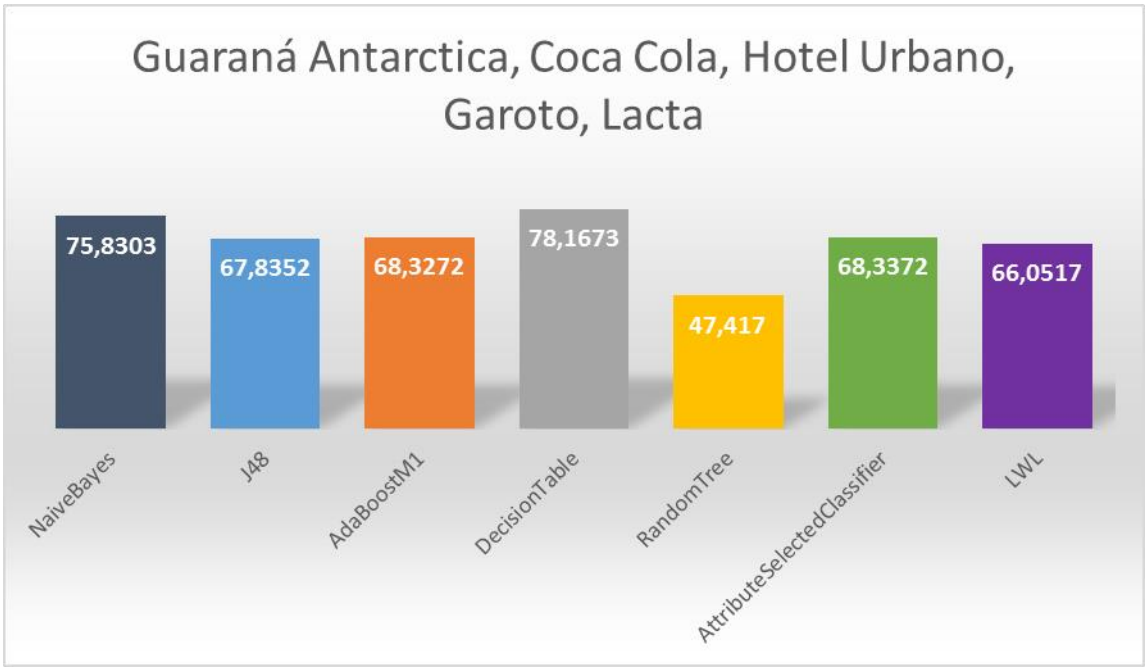


Figura 28 - Resultados da Classificação do Fator de Impacto por Algoritmo

5.1 Atributos que exercem influência no fator de impacto do corpus Guaraná Antarctica.

Nesta tarefa foram analisados os seguintes elementos com o objetivo de identificar os fatores que exercem influência no fator de impacto proposto no presente trabalho, dentre eles são listados os seguintes:

1. Número de palavras por Fator de Impacto;
2. Número de curtidas, comentários e compartilhamento por Fator de Impacto;
3. Número de Unigram por Fator de Impacto;
4. Número de Bigram por Fator de Impacto;
5. Número de Trigram por Fator de Impacto;
6. Número de Ngram por Fator de Impacto;
7. Número de mensagens em cada mês por Fator de Impacto;
8. Número de mensagens em cada dia da semana por Fator de Impacto;
9. Número de mensagens em cada turno por Fator de Impacto;
10. Número de tipos de mensagens por Fator de Impacto.

A ordem apresentada foi considerada pela sequencia que estes dez atributos obtidos através do uso de técnicas de PLN bem com DCBD aparecem na base de dados. A intenção é identificar quais atributos contribuem para identificar as notícias com maior fator de impacto ou seja com maior número de interações, e através destes atributos propor um modelo de predição de interações em redes sociais.

5.1.1 Número de palavras por Fator de Impacto do Corpus Guaraná Antarctica

Na sequência é apresentada a forma que foi contabilizada o total de palavras presentes nas notícias publicadas na página de fãs Guaraná Antarctica classificadas como fator de impacto alto, médio e baixo. Na figura 29 evidenciamos os resultados encontrados e também uma análise dos resultados.

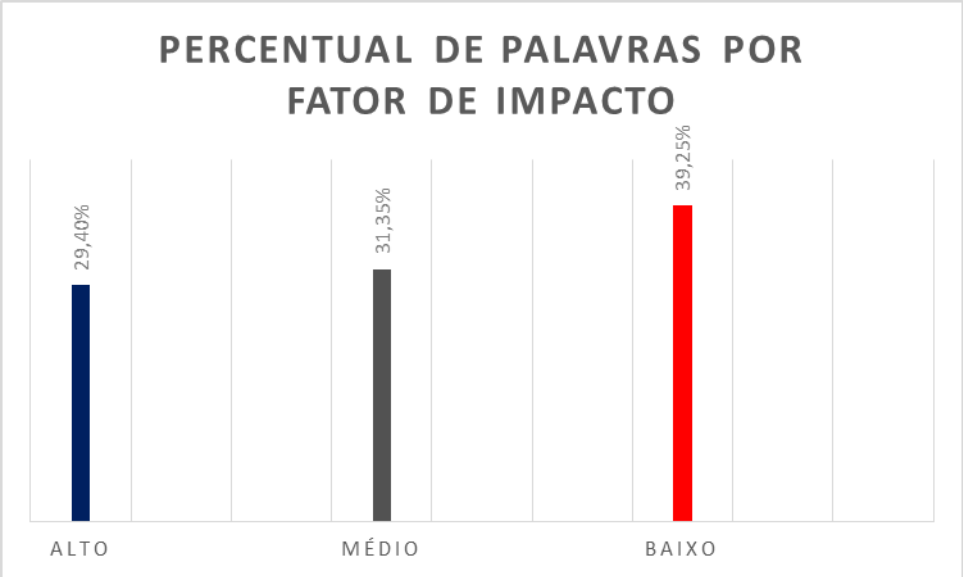


Figura 29 – Palavras por Fator de Impacto do Corpus *Guaraná Antarctica*

Conforme os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 29,40% de número de palavras, 31,35% do número de palavras classificadas como Fator de Impacto Médio e 39,25% do número de palavras classificadas como Fator de Impacto Baixo. Ou seja, quanto maior o número de palavras presentes menor será o fator de Impacto.

5.1.2 Número de curtidas, comentários e compartilhamento por Fator de Impacto do Corpus Guaraná Antarctica.

Na sequência é apresentada a forma que foi contabilizada as seguintes interações: número de curtidas, comentários e compartilhamentos presentes nas notícias publicadas na página de fãs Guaraná Antarctica classificadas como fator de impacto alto, médio e baixo. Na figura 30 evidenciamos os resultados e também uma análise dos resultados encontrados.

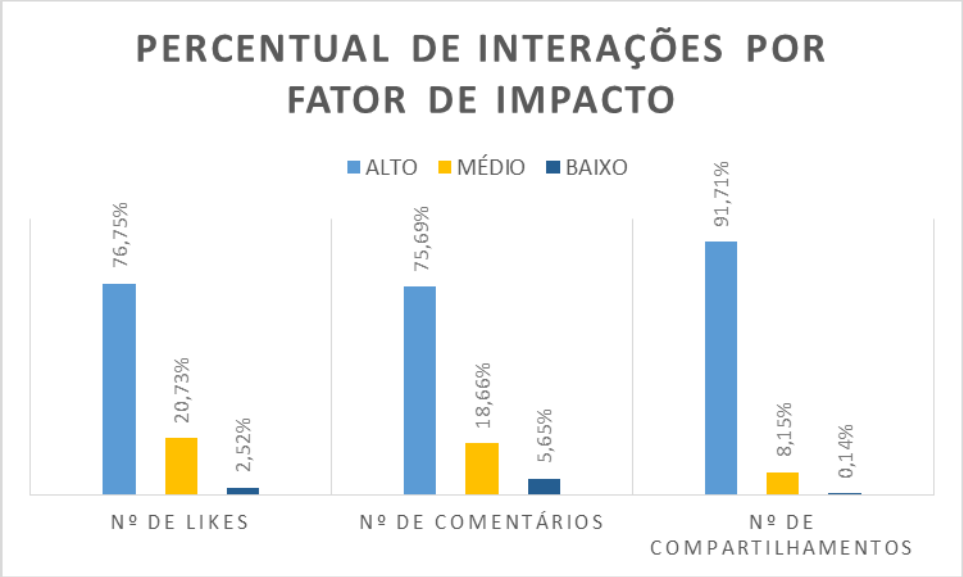


Figura 30 – Interações por Fator de Impacto do Corpus *Guaraná Antarctica*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 76,75% do número de likes, 75,69% do número de comentários e 91,71% do número de compartilhamentos.

De outra forma significa que no corpus de notícias extraídas da página de fãs do Guaraná Antárctica quanto maior o número de likes, comentários e compartilhamentos presentes nas notícias divulgadas na página de fãs da rede social Facebook maior será seu Fator de Impacto.

Observa-se também que o número de compartilhamentos supera o percentual de likes e comentários em tais notícias. Umas das explicações desse resultado é devido a classe fator de impacto ser composta pela média das três interações.

5.1.3 Número de Unigram por Fator de Impacto do Corpus Guaraná Antarctica

Na sequência é apresentada a forma que foi contabilizada as palavras mais frequentes presentes nas notícias publicadas na página de fãs Guaraná Antarctica classificadas como fator de impacto alto, médio e baixo. Na figura 31 evidenciamos os resultados e também uma análise dos resultados encontrados.

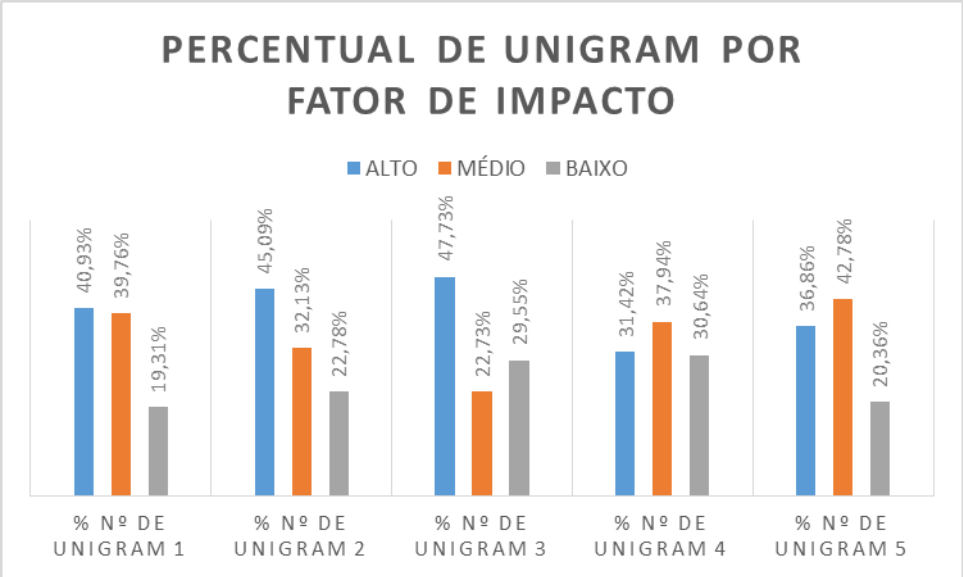


Figura 31 - Percentual de Unigram por Fator de Impacto do Corpus *Guaraná Antarctica*

Segundo os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 40,93% de palavras mais frequentes, 45,09% com a segunda palavra mais frequente, 47,73% com a terceira palavra mais frequente, 31,42 com a quarta palavra mais frequente e 36,86% com a quinta palavra mais frequente.

5.1.4 Número de Bigram por Fator de Impacto do Corpus Guaraná Antarctica

Na sequência é apresentada a forma que foi contabilizada os pares de palavras mais frequentes presentes nas notícias publicadas na página de fãs Guaraná Antarctica classificadas como fator de impacto alto, médio e baixo. Na figura 32 evidenciamos os resultados e também uma análise dos resultados encontrados.

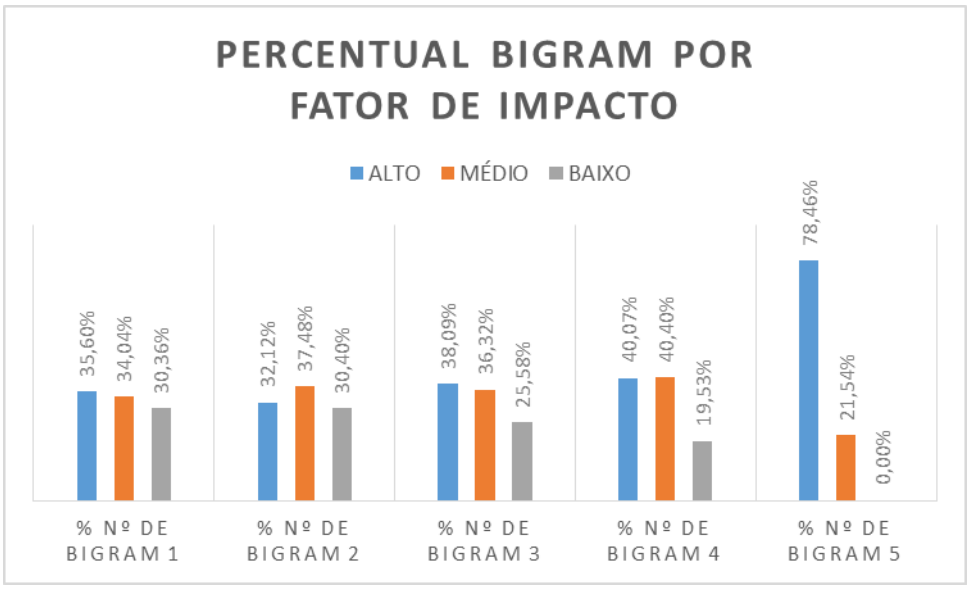


Figura 32 - Percentual de Bigram por Fator de Impacto do Corpus *Guaraná Antarctica*

Consoante os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 35,60% primeiro conjunto de duas palavras com maior frequência, 32,12% segundo conjunto de duas palavras com maior frequência, 38,09% terceiro conjunto de duas palavras com maior frequência, 40,07% quarto conjunto de duas palavras com maior frequência e 78,46% quinta palavra com maior frequência.

5.1.5 Número de Trigram por Fator de Impacto do Corpus Guaraná Antarctica

Na sequência é apresentada a forma que foi contabilizada os conjuntos de três palavras mais frequentes presentes nas notícias publicadas na página de fãs Guaraná Antarctica classificadas como fator de impacto alto, médio e baixo. Na figura 33 evidenciamos os resultados e também uma análise dos resultados encontrados.

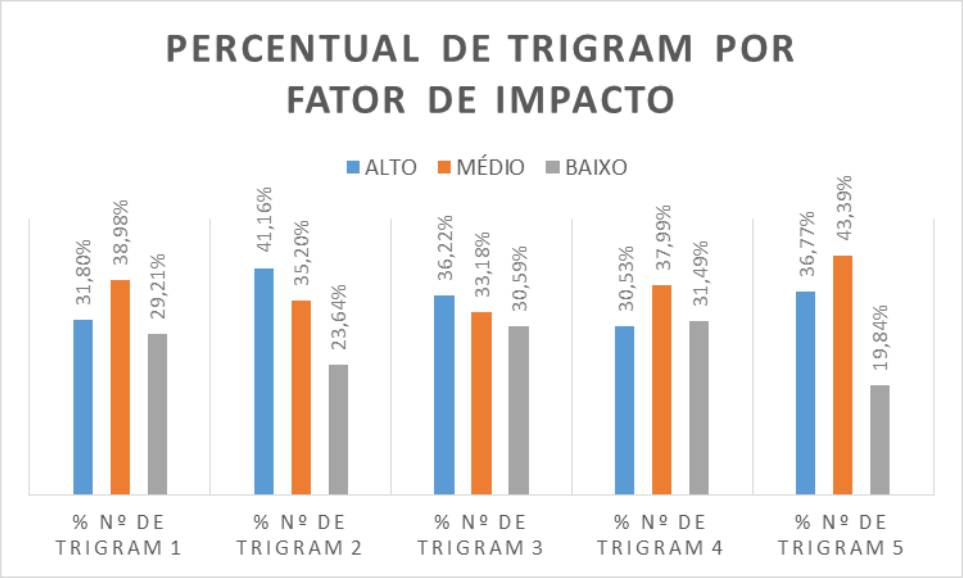


Figura 33 - Percentual de Trigram por Fator de Impacto do Corpus *Guaraná Antarctica*

De acordo com os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, o maior resultado para esta classe é observado para o 2º conjunto de três palavras com 41,16% e em segundo lugar para o quinto conjunto de três palavras com 36,77%. Em geral os conjuntos formados com três palavras mais frequentes utilizando a técnica trigram não dizem muito sobre uma perspectiva de influência, uma vez que os resultados para ambas as classes ficam próximos e com um percentual abaixo de 50%.

5.1.6 Número de Ngram por Fator de Impacto do Corpus Guaraná Antarctica

Na sequência é apresentada a forma que foi contabilizada os conjuntos de cinco e dez palavras mais frequentes presentes nas notícias publicadas na página de fãs Guaraná Antarctica classificadas como fator de impacto alto, médio e baixo. Na figura 34 evidenciamos os resultados e também uma análise dos resultados encontrados.

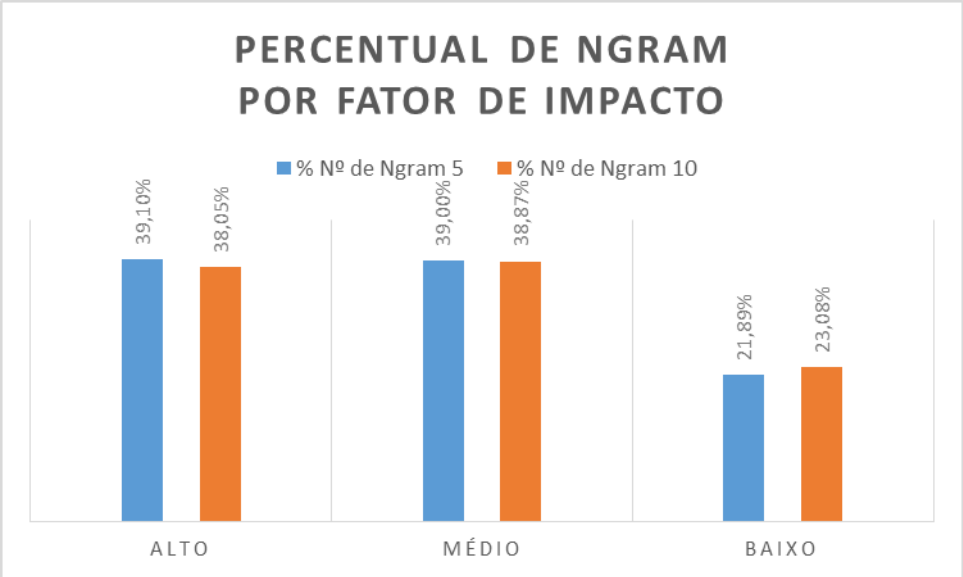


Figura 34 - Percentual de Ngram por Fator de Impacto do Corpus *Guaraná Antarctica*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 39,10% para p primeiro conjunto de cinco palavras com maior frequência, às classificadas como Fator de Impacto Médio 39,00% e as classificadas como Fator de Impacto Baixo 23,08% considerando o conjunto de dez palavras com maior frequência.

De maneira geral é observado que o percentual de NGRAM usando as cinco palavras mais frequentes ficou com percentual acima para as notícias classificadas como fator de impacto alto e médio. E que o NGRAM usando as dez palavras mais frequentes apresenta um percentual de 23,08% para notícias classificadas como fator de impacto baixo.

5.1.7 Número de mensagens em meses por Fator de Impacto do Corpus Guaraná Antarctica

Na sequência é apresentada a forma que foi contabilizada os conjuntos de mensagens por mês presentes nas notícias publicadas na página de fãs Guaraná Antarctica classificadas como fator de impacto alto, médio e baixo. Na figura 35 evidenciamos os resultados e também uma análise dos resultados encontrados.

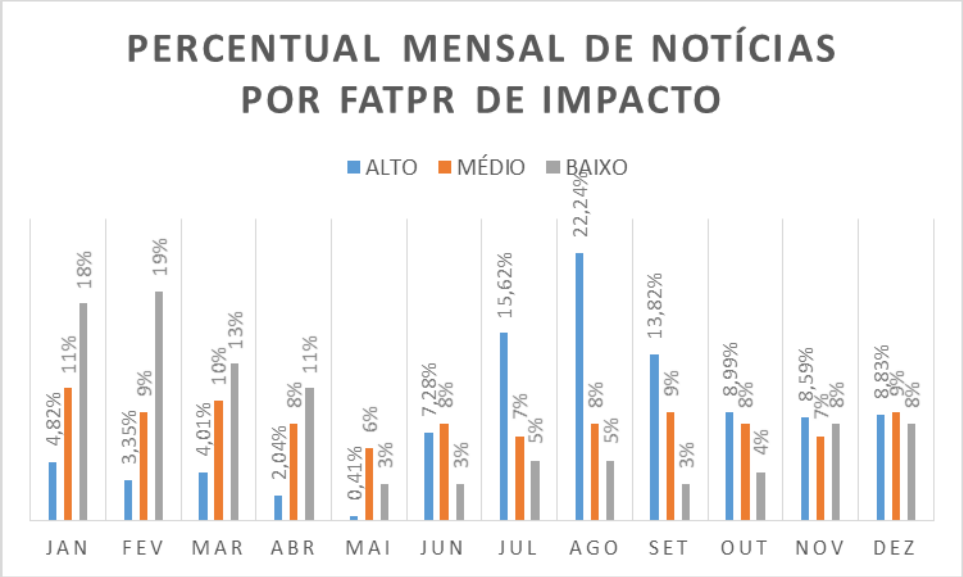


Figura 35 - Percentual de Notícias por mês e Fator de Impacto do Corpus *Guaraná Antarctica*

Consoante os resultados apresentados acima pode-se perceber que se tratando de notícias divulgadas por mês presente na página de fãs do Guaraná Antarctica, o atributo que aparece com maior intensidade é o mês agosto em notícias classificadas como fator de impacto alto. Entretanto já em notícias classificadas como fator de impacto médio e baixo é evidenciado os meses de janeiro e fevereiro.

5.1.8 Número de mensagens em dias da semana por Fator de Impacto do Corpus Guaraná Antarctica

Na sequência é apresentada a forma que foi contabilizada os conjuntos de mensagens por dia da semana presentes nas notícias publicadas na página de fãs Guaraná Antarctica classificadas como fator de impacto alto, médio e baixo. Na figura 36 evidenciamos os resultados e também uma análise dos resultados encontrados.

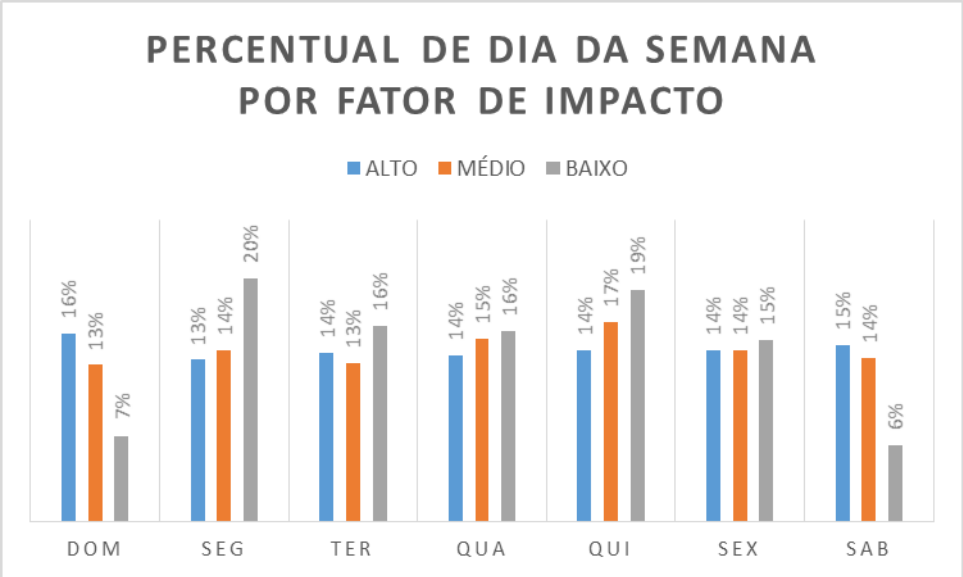


Figura 36 - Percentual de Notícias por dia da semana e Fator de Impacto do Corpus *Guaraná Antarctica*

De acordo com os resultados apresentados acima pode-se perceber que se tratando de notícias divulgadas por dia da semana presente na página de fãs do Guaraná Antarctica, o atributo que aparece com maior intensidade é o domingo em notícias classificadas como fator de impacto alto. Entretanto já em notícias classificadas como fator de impacto médio e baixo é evidenciado os dias da semana quinta e segunda.

5.1.9 Número de mensagens em turnos por Fator de Impacto do Corpus Guaraná Antarctica

Na sequência é apresentada a forma que foi contabilizada as mensagens por turno presentes nas notícias publicadas na página de fãs Guaraná Antarctica classificadas como fator de impacto alto, médio e baixo. Na figura 37 evidenciamos os resultados e também uma análise dos resultados encontrados.

Esta tarefa tem o objetivo de identificar qual o turno que possui maior incidência nas mensagens classificadas como fator de impacto alto, médio e baixo.

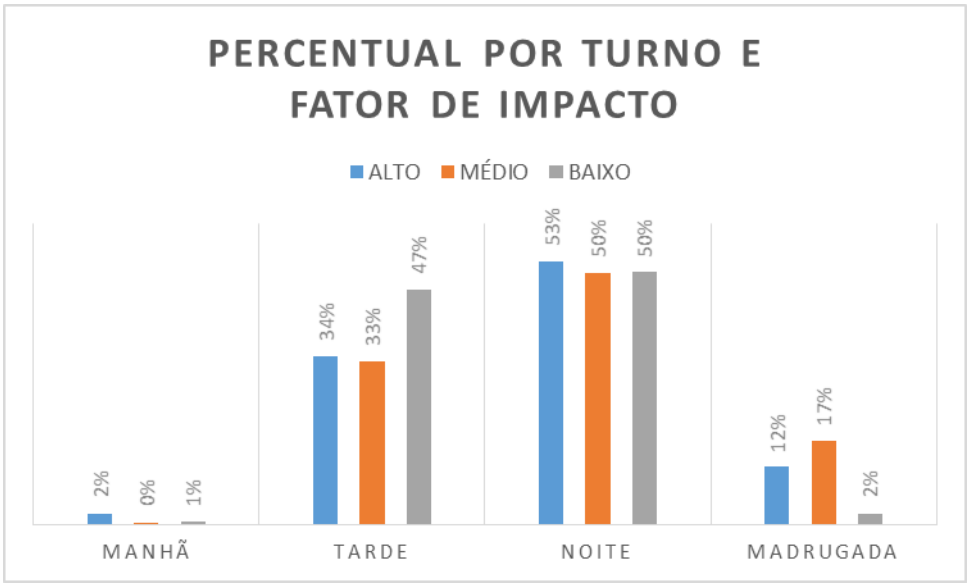


Figura 37 - Percentual de Notícias por turno e Fator de Impacto do Corpus *Guaraná Antarctica*

Conforme os resultados apresentados acima pode-se perceber que se tratando de notícias divulgadas em turnos presente na página de fãs do Guaraná Antarctica, o atributo que aparece com maior intensidade é o turno da noite seguida do turno da tarde em notícias classificadas como fator de impacto alto, médio e baixo.

5.1.10 Número de tipos de mensagens por Fator de Impacto do Corpus Guaraná Antarctica

Na sequência é apresentada a forma que foi contabilizada os tipos de mensagens presentes nas notícias publicadas na página de fãs Guaraná Antarctica classificadas como fator de impacto alto, médio e baixo. Na figura 38 evidenciamos os resultados e também uma análise dos resultados encontrados.

Esta tarefa tem o objetivo de identificar qual o tipo de publicação possui maior incidência nas mensagens classificadas como fator de impacto alto, médio e baixo.

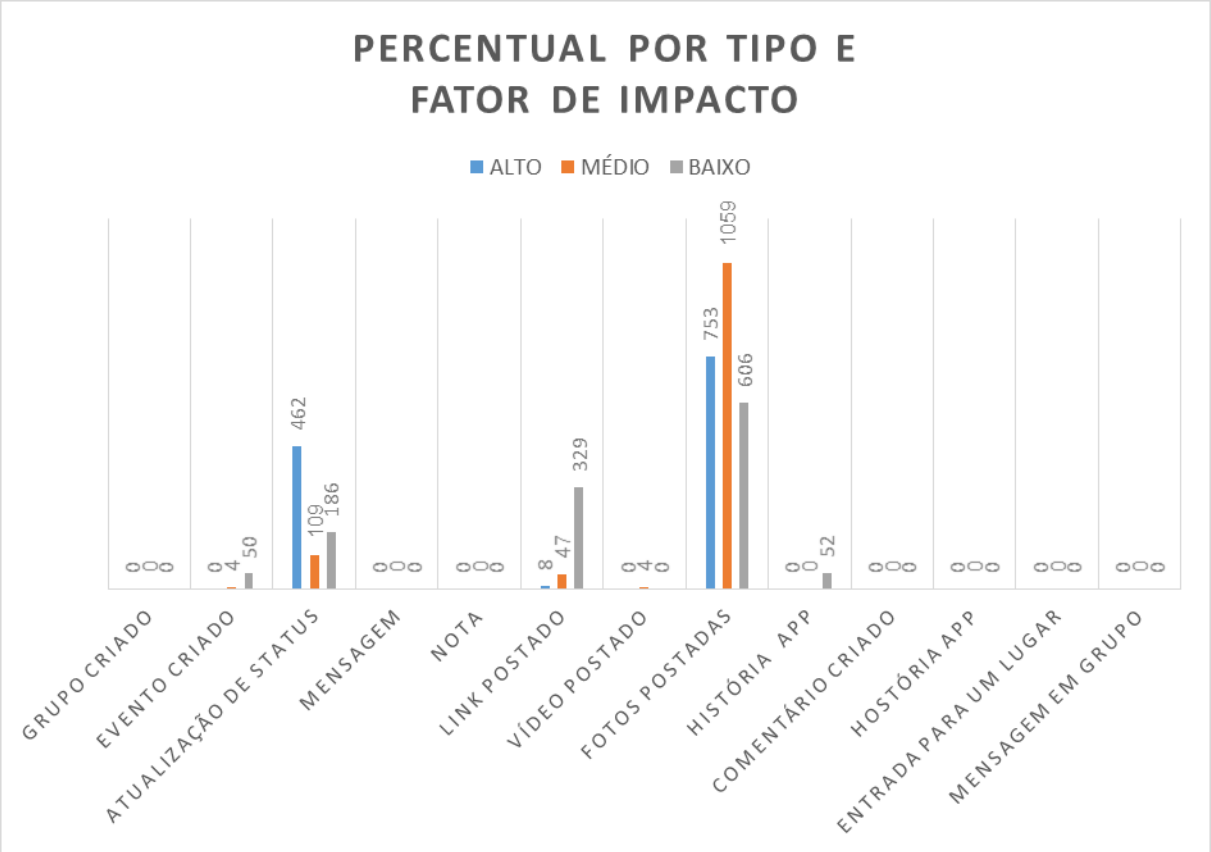


Figura 38 - Percentual de Notícias por Tipo e Fator de Impacto do Corpus *Guaraná Antarctica*

Através dos resultados apresentados acima pode-se perceber que se tratando de tipos de notícias presente na página de fãs do Guaraná Antarctica, o tipo foto postado, aparece com maior intensidade em notícias classificadas como fator de impacto alto, médio e baixo. Contudo em notícias com fator impacto alto é observado o tipo atualização de status. Também em notícias divulgadas classificadas como fator de impacto baixo é verificado o tipo link postado.

5.2 Atributos que exercem influência no fator de impacto do corpus Coca Cola.

Da mesma forma apresentada acima, nesta tarefa foram analisados os seguintes elementos com o objetivo de identificar os fatores que exercem influência no fator de impacto proposto no presente trabalho, dentre eles são listados os seguintes:

1. Número de palavras por Fator de Impacto;
2. Número de curtidas, comentários e compartilhamento por Fator de Impacto;
3. Número de Unigram por Fator de Impacto;
4. Número de Bigram por Fator de Impacto;
5. Número de Trigram por Fator de Impacto;
6. Número de Ngram por Fator de Impacto;
7. Número de mensagens em cada mês por Fator de Impacto;
8. Número de mensagens em cada dia da semana por Fator de Impacto;
9. Número de mensagens em cada turno por Fator de Impacto;
10. Número de tipos de mensagens por Fator de Impacto.

5.2.1 Número de palavras por Fator de Impacto do Corpus Coca Cola

Na figura 39 é apresentada a contabilização total de palavras presentes nas notícias publicadas na página de fãs da Coca Cola classificada como fator de impacto alto, médio e baixo.

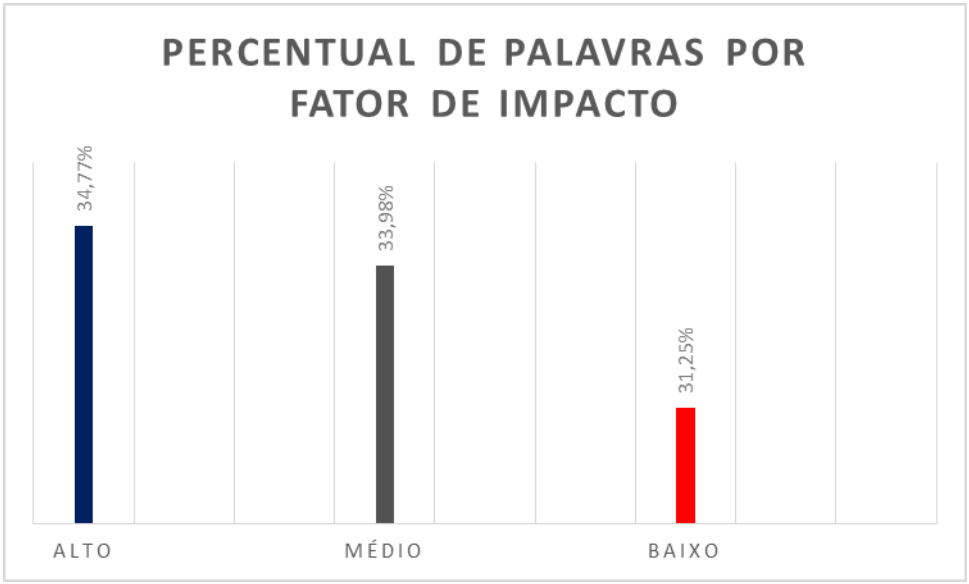


Figura 39 - Percentual de Palavras por Fator de Impacto do Corpus *Coca Cola*

Segundo os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 34,77% de número de palavras, 33,98% do número de palavras classificadas como Fator de Impacto Médio e 31,25% do número de palavras classificadas como Fator de Impacto Baixo.

5.2.2 Número de curtidas, comentários e compartilhamento por Fator de Impacto do Corpus Coca Cola.

Na figura 40 é apresentada a forma que foi contabilizada as seguintes interações: número de curtidas, comentários e compartilhamentos presentes nas notícias publicadas na página de fãs Coca Cola classificado como fator de impacto alto, médio e baixo.

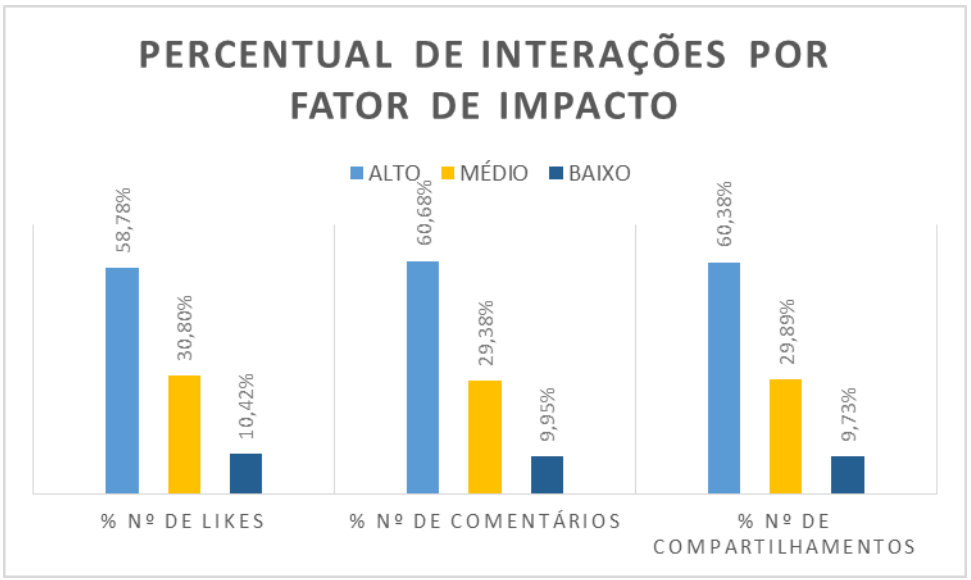


Figura 40 – Percentual de Interações por Fator de Impacto do Corpus *Coca Cola*

Consoante os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 58,78% do número de likes, 60,68% do número de comentários e 60,38% do número de compartilhamentos.

De outra forma significa que no corpus de notícias extraídas da página de fãs da Coca Cola quanto maior o número de likes, comentários e compartilhamentos presentes nas notícias divulgadas na página de fãs da rede social Facebook maior será seu Fator de Impacto.

5.2.3 Número de Unigram por Fator de Impacto do Corpus Coca Cola

Na figura 41 é apresentada a forma que foi contabilizada as palavras mais frequentes presentes nas notícias publicadas na página de fãs Guaraná Coca Colam classificadas como fator de impacto alto, médio e baixo.

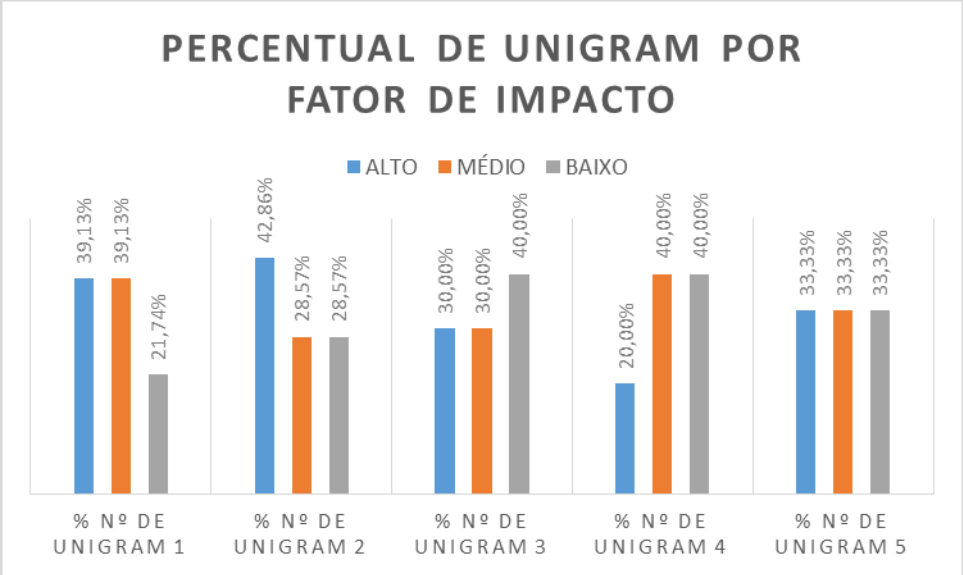


Figura 41 - Percentual de Unigram por Fator de Impacto do Corpus *Coca Cola*

De acordo com os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 39,13% de palavras mais frequentes, 42,86% com a segunda palavra mais frequente, 30,00% com a terceira palavra mais frequente, 20,00 com a quarta palavra mais frequente e 33,33% com a quinta palavra mais frequente.

É observado também que o melhor resultado foi com a segunda palavras mais frequente com 42,86% para a classe fator de impacto alto seguido de 40% utilizando a terceira e quarta palavras mais frequentes para as classes fator de impacto média e baixa.

5.2.4 Número de Bigram por Fator de Impacto do Corpus Coca Cola

Na figura 42 é apresentada a forma que foi contabilizada os pares de palavras mais frequentes presentes nas notícias publicadas na página de fãs Coca Colam classificados como fator de impacto alto, médio e baixo.

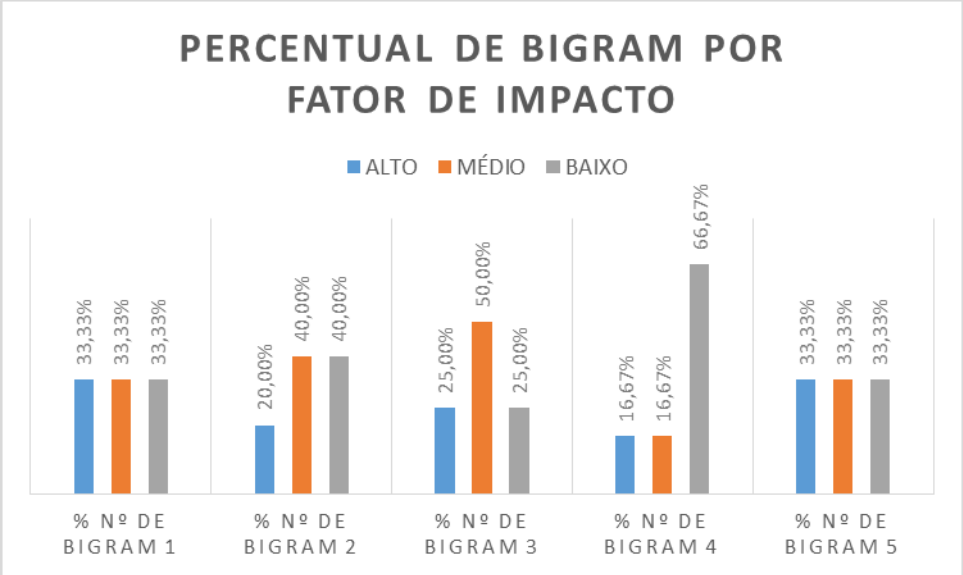


Figura 42 - Percentual de Bigram por Fator de Impacto do Corpus *Coca Cola*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 33,33% primeiro conjunto de duas palavras com maior frequência, 20,00% segundo conjunto de duas com maior frequência, 25,00% terceiro conjunto de duas com maior frequência, 16,00% quarto conjunto de duas palavras com maior frequência e 33,33% quinta conjunto de duas palavras com maior frequência.

5.2.5 Número de Trigram por Fator de Impacto do Corpus Coca Cola

Na figura 43 é apresentada a forma que foi contabilizada os conjuntos de três palavras mais frequentes presentes nas notícias publicadas na página de fãs Coca Colam classificados como fator de impacto alto, médio e baixo.

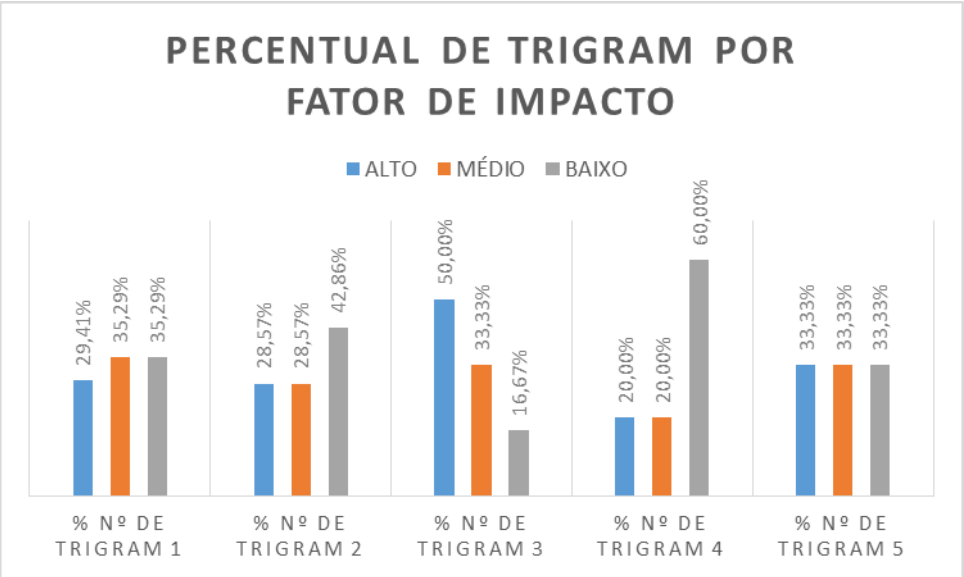


Figura 43 - Percentual de Trigram por Fator de Impacto do Corpus *Coca Cola*

Conforme os resultados apresentados pode-se perceber que em notícias publicadas, classificadas como Fator de Impacto Alto, o terceiro conjunto de três palavras conseguiram classificar com percentual de 50% notícias com fator de impacto alto. Além disso é observado que o quarto conjunto de palavras mais frequentes obtém um percentual de 60% em notícias classificadas como fator de impacto baixo.

5.2.6 Número de Ngram por Fator de Impacto do Corpus Coca Cola

Na figura 44 é apresentada a forma que foi contabilizada os conjuntos de cinco e dez palavras mais frequentes presentes nas notícias publicadas na página de fãs Coca Colam classificados como fator de impacto alto, médio e baixo.

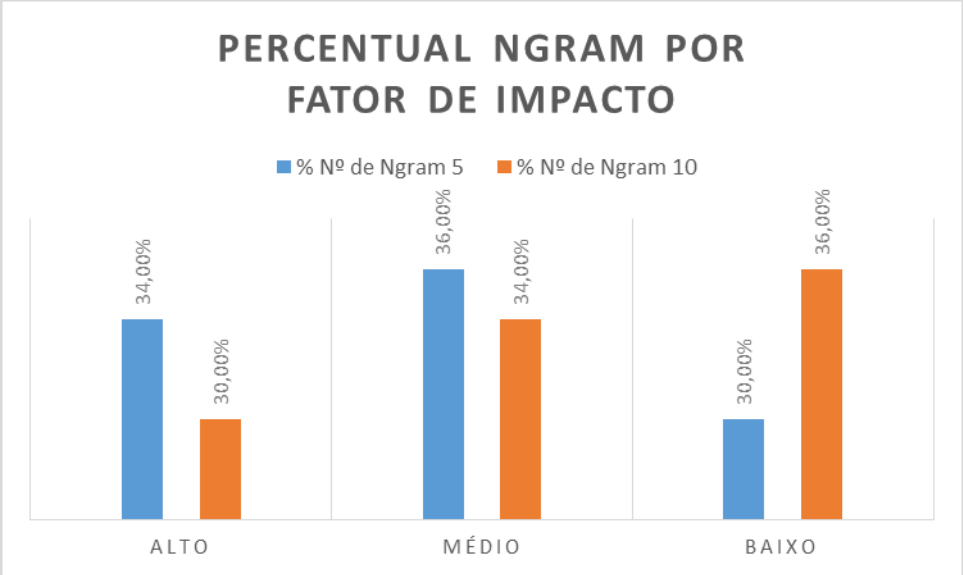


Figura 44 - Percentual de Ngram por Fator de Impacto do Corpus *Coca Cola*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 34,00% primeiro conjunto de cinco palavras com maior frequência, as classificadas como Fator de Impacto Médio 36,00% segundo conjunto de cinco palavras com maior frequência e as classificadas como Fator de Impacto Baixo 30,00% terceiro conjunto de cinco palavras com maior frequência.

5.2.7 Número de mensagens em meses por Fator de Impacto do Corpus Coca Cola

Na figura 45 é apresentada a forma que foi contabilizada as mensagens mais frequentes por meses presentes nas notícias publicadas na página de fãs Coca Colam classificadas como fator de impacto alto, médio e baixo.

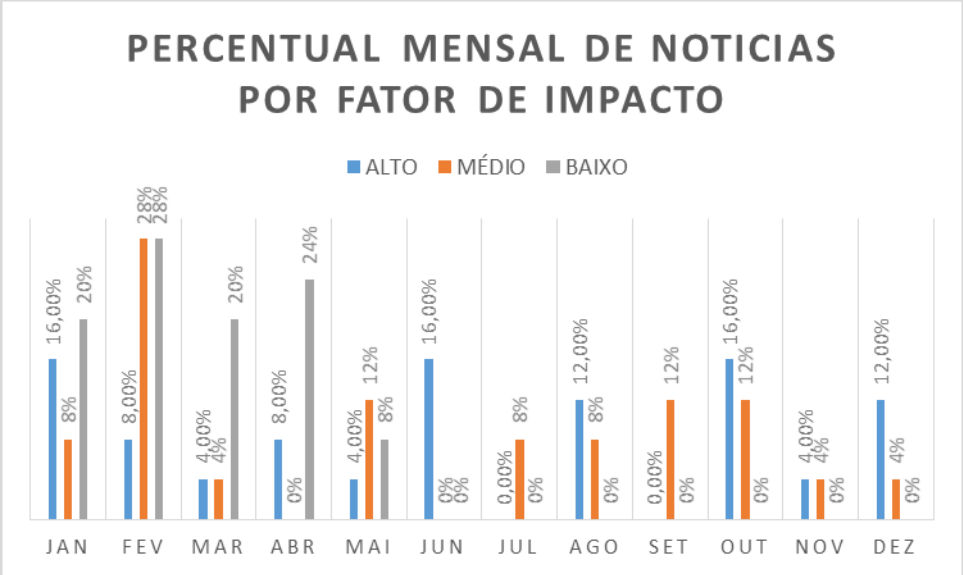


Figura 45 - Percentual de Notícias por mês e Fator de Impacto do Corpus *Coca Cola*

Segundo os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por meses presente na página de fãs da Coca Cola, revela que em notícias classificadas como fator de impacto alto os meses com mais mensagens foram janeiro, julho e outubro. No entanto em notícias classificadas cm fator de impacto médio e baixo é evidenciado o mês de fevereiro.

5.2.8 Número de mensagens em cada dia da semana por Fator de Impacto do Corpus Coca Cola

Na figura 46 é apresentada a forma que foi contabilizada a frequência de mensagens por dia da semana presentes nas notícias publicadas na página de fãs Coca Cola classificada como fator de impacto alto, médio e baixo.

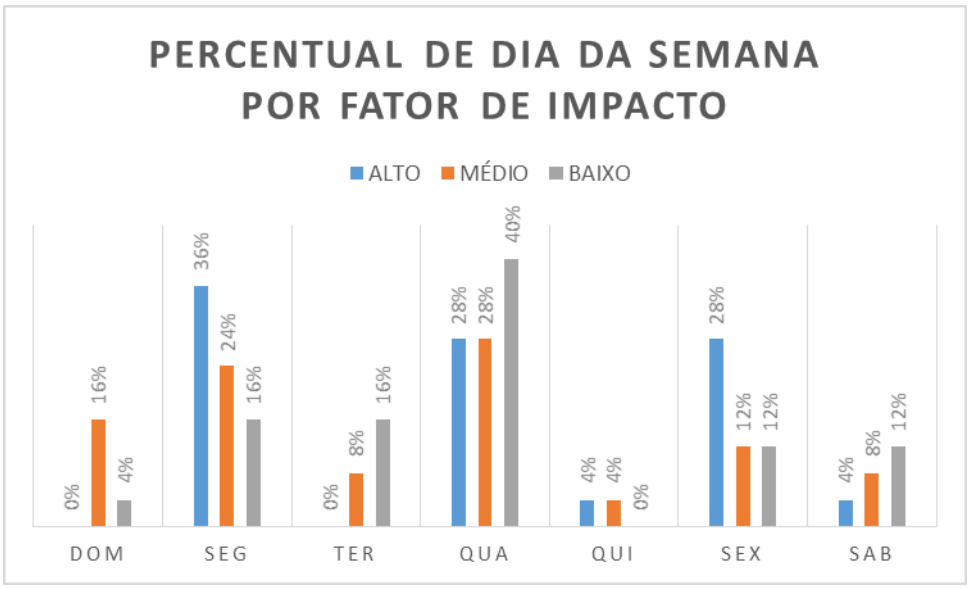


Figura 46 - Percentual de Notícias por dia da semana e Fator de Impacto do Corpus *Coca Cola*

Consoante os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por dia da semana presente na página de fãs da Coca Cola, revela que em notícias classificadas como fator de impacto alto o dia da semana vencedor foi segunda. Contudo em notícias classificadas como fator de impacto médio e baixo o dia da semana foi quarta.

5.2.9 Número de mensagens em turnos por Fator de Impacto do Corpus Coca Cola

Na figura 47 é apresentada a forma que foi contabilizada a frequência de mensagens por turno presentes nas notícias publicadas na página de fãs Coca Cola classificada como fator de impacto alto, médio e baixo.

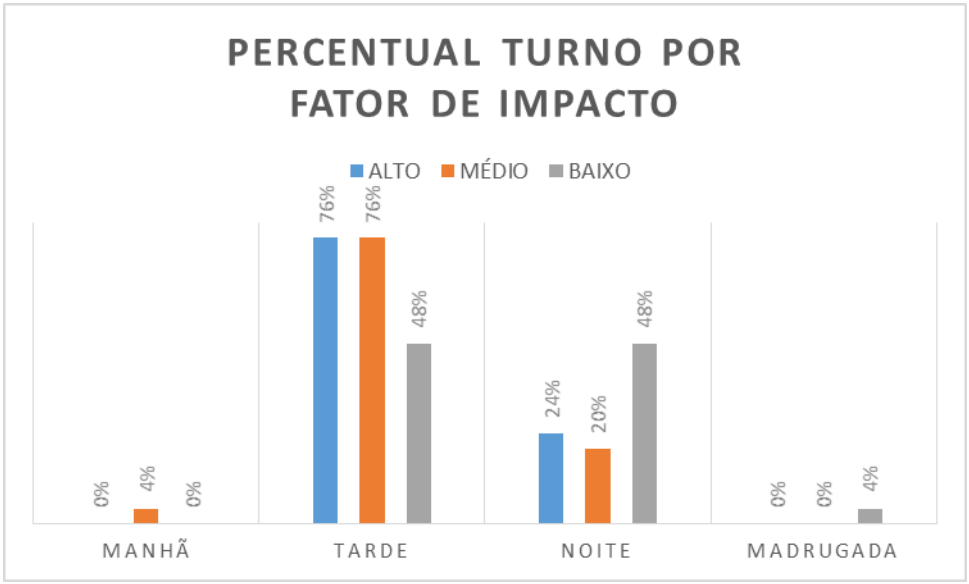


Figura 47 - Percentual do número de mensagens publicadas em turnos por Fator de Impacto do Corpus *Coca Cola*

De acordo com os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por turno presente na página de fãs da Coca Cola, revela que tanto em notícias classificadas como fator de impacto alto, médio e baixo o turno da tarde concentra uma maior quantidade de notícias, e em segundo lugar também do turno da noite.

5.2.10 Número de tipos de mensagens por Fator de Impacto do Corpus Coca Cola

Na figura 48 é apresentada a forma que foi contabilizada a frequência de tipos de mensagens presentes nas notícias publicadas na página de fãs Coca Cola classificada como fator de impacto alto, médio e baixo.

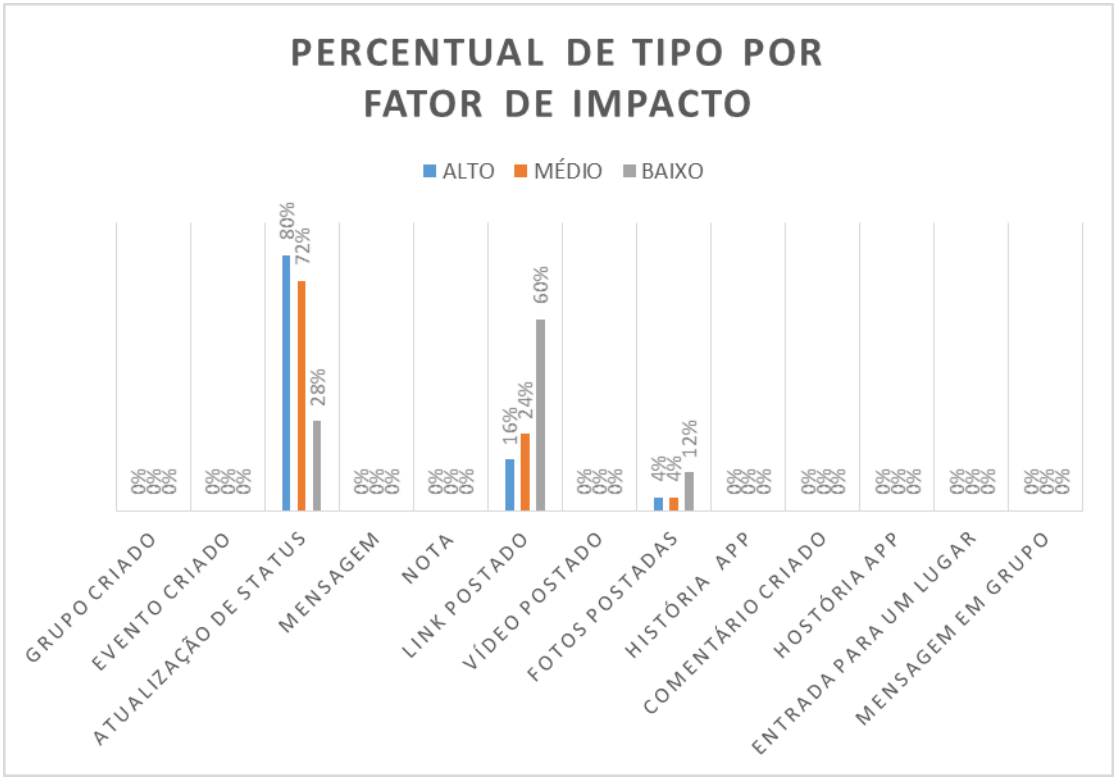


Figura 48 - Percentual de mensagens publicadas em tipos por Fator de Impacto do Corpus *Coca Cola*

Através dos resultados apresentados acima pode-se perceber que se tratando de notícias classificadas como fator de impacto alto e médio o tipo de notícia que mais ocorre é de Atualização de Sataus. E em notícias classificadas como fator de impacto baixo o tipo evidenciado foi o de link postado.

5.3 Atributos que Exercem Influência no Fator de Impacto do corpus Hotel Urbano.

Da mesma forma, nesta tarefa foram analisados os seguintes elementos com o objetivo de identificar os fatores que exercem influência no fator de impacto proposto no presente trabalho, dentre eles são listados os seguintes:

1. Número de palavras por Fator de Impacto;
2. Número de curtidas, comentários e compartilhamento por Fator de Impacto;
3. Número de Unigram por Fator de Impacto;
4. Número de Bigram por Fator de Impacto;
5. Número de Trigram por Fator de Impacto;
6. Número de Ngram por Fator de Impacto;
7. Número de mensagens em cada mês por Fator de Impacto;
8. Número de mensagens em cada dia da semana por Fator de Impacto;
9. Número de mensagens em cada turno por Fator de Impacto;
10. Número de tipos de mensagens por Fator de Impacto.

5.3.1 Número de palavras por Fator de Impacto do Corpus Hotel Urbano

Na figura 49 é apresentada a contabilização total de palavras presentes nas notícias publicadas na página de fãs do Hotel Urbano classificadas como fator de impacto alto, médio e baixo.

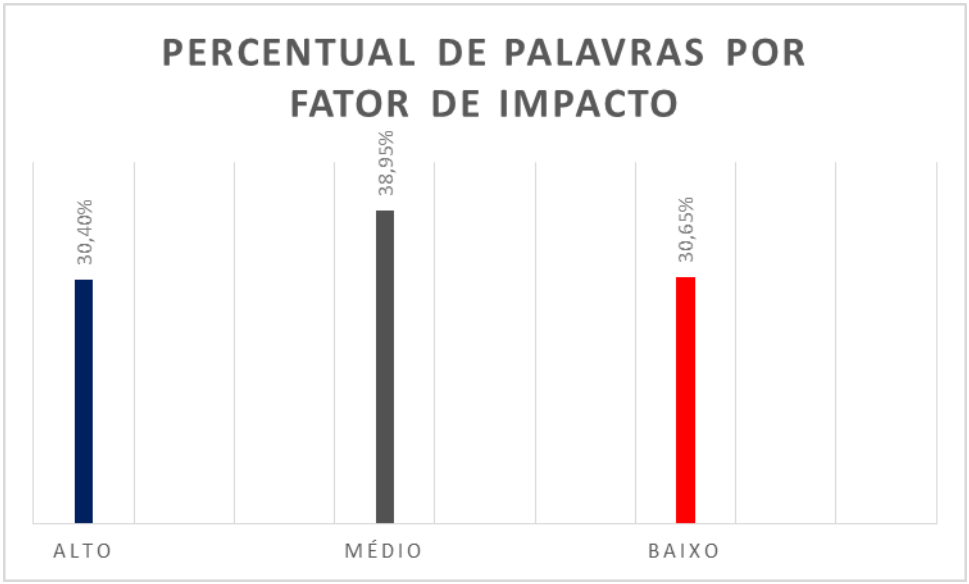


Figura 49 - Percentual de Palavras por Fator de Impacto do Corpus *Hotel Urbano*

Conforme os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 30,40% de número de palavras, 38,95% do número de palavras classificadas como Fator de Impacto Médio e 30,65% do número de palavras classificadas como Fator de Impacto Baixo.

5.3.2 Número de curtidas, comentários e compartilhamento por Fator de Impacto do Corpus Hotel Urbano

Na figura 50 é apresentada a forma que foi contabilizada as seguintes interações: número de curtidas, comentários e compartilhamentos presentes nas notícias publicadas na página de fãs Hotel Urbano classificado como fator de impacto alto, médio e baixo.

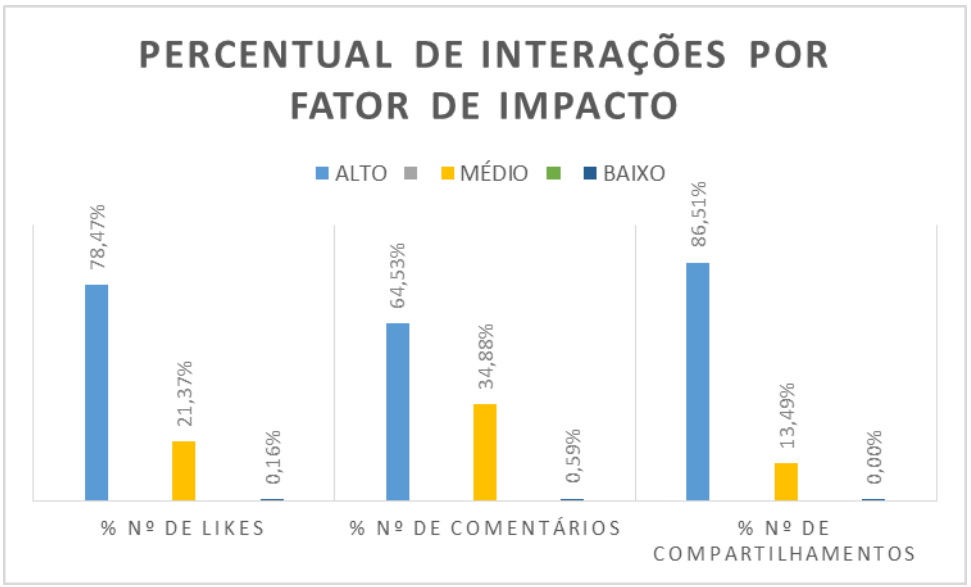


Figura 50 – Percentual de Interações por Fator de Impacto do Corpus *Hotel Urbano*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 78,47% do número de likes, 64,53% do número de comentários e 86,51% do número de compartilhamentos.

De outra forma significa que no corpus de notícias extraídas da página de fãs do Hotel Urbano quanto maior o número de likes, comentários e compartilhamentos presentes nas notícias divulgadas na página de fãs da rede social Facebook maior será seu Fator de Impacto.

Observa-se também que o número de compartilhamentos supera o percentual de likes e comentários em tais notícias.

5.3.3 Número de Unigram por Fator de Impacto do Corpus Hotel Urbano

Na figura 51 é apresentada a forma que foi contabilizada as palavras mais frequentes presentes nas notícias publicadas na página de fãs Hotel urbano classificado como fator de impacto alto, médio e baixo.

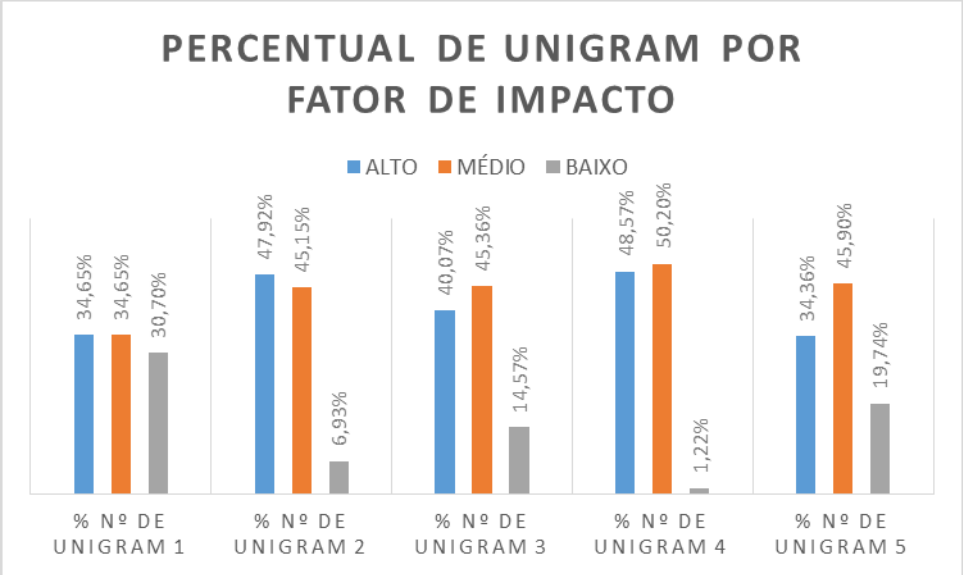


Figura 51 - Percentual de Unigram por Fator de Impacto do Corpus *Hotel Urbano*

Segundo os resultados apresentados é observado que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 34,65% de palavras mais frequentes, 47,92% com a segunda palavra mais frequente, 40,07% com a terceira palavra mais frequente, 48,57% com a quarta palavra mais frequente e 34,36% com a quinta palavra mais frequente.

5.3.4 Número de Bigram por Fator de Impacto do Corpus Hotel Urbano

Na figura 52 é apresentada a forma que foi contabilizada os pares de palavras mais frequentes presentes nas notícias publicadas na página de fãs Hotel Urbano classificado como fator de impacto alto, médio e baixo.

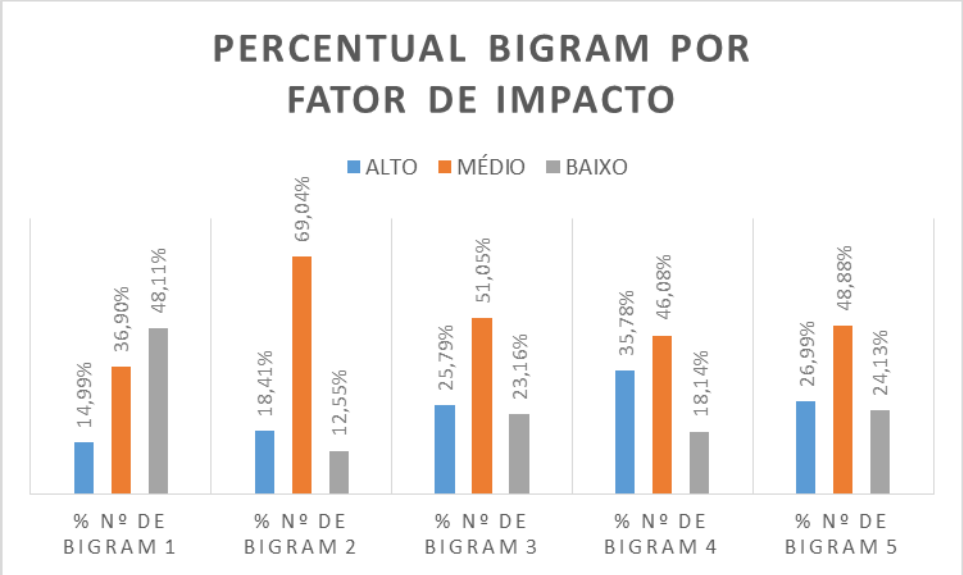


Figura 52 - Percentual de Bigram por Fator de Impacto do Corpus *Hotel Urbano*

Consoante os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 14,99% primeiro conjunto de duas palavras com maior frequência, 18,41% segundo conjunto de duas com maior frequência, 25,79% terceiro conjunto de duas com maior frequência, 35,78% quarto conjunto de duas palavras com maior frequência e 26,99% quinta conjunto de duas palavras com maior frequência.

5.3.5 Número de Trigram por Fator de Impacto do Corpus Hotel Urbano

Na figura 53 é apresentada a forma que foi contabilizada os conjuntos de três palavras mais frequentes presentes nas notícias publicadas na página de fãs Hotel Urbano classificado como fator de impacto alto, médio e baixo.

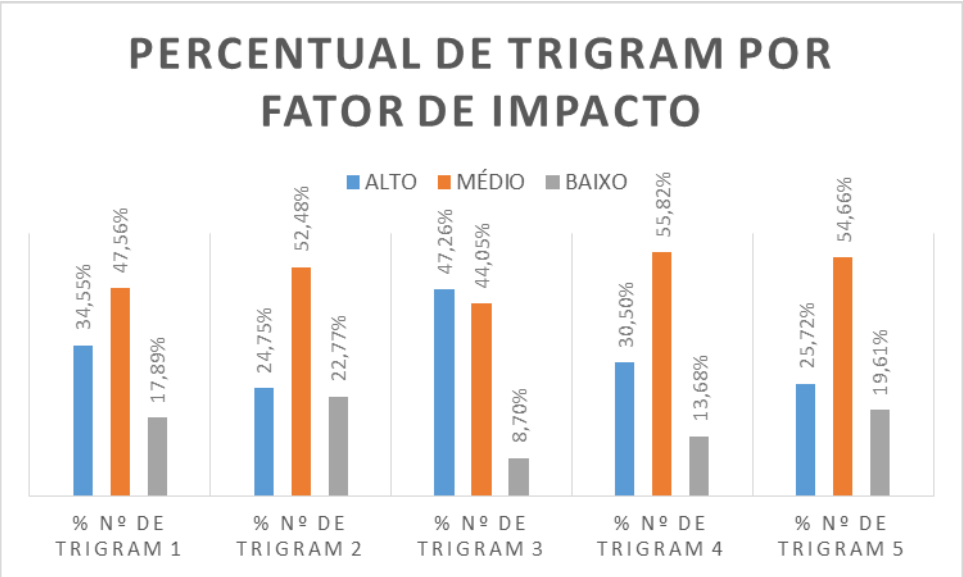


Figura 53- Percentual de Trigram por Fator de Impacto do Corpus *Hotel Urbano*

De acordo com os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 34,55% primeiro conjunto de três palavras com maior frequência, 24,75% segundo conjunto de três com maior frequência, 47,26% terceiro conjunto de três palavras com maior frequência, 30,50% quarto conjunto de três palavras com maior frequência e 25,72% quinto conjunto de três palavras com maior frequência.

5.3.6 Número de Ngram por Fator de Impacto do Corpus Hotel Urbano

Na figura 54 é apresentada a forma que foi contabilizada os conjuntos de cinco e dez palavras mais frequentes presentes nas notícias publicadas na página de fãs Hotel urbano classificado como fator de impacto alto, médio e baixo.

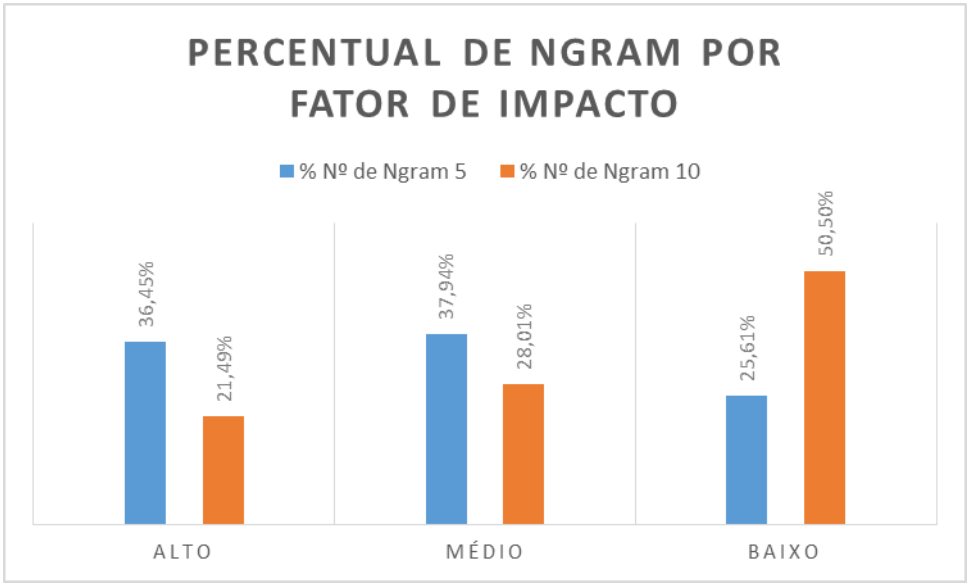


Figura 54 - Percentual de Ngram por Fator de Impacto do Corpus *Hotel Urbano*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 36,45% primeiro conjunto de cinco palavras com maior frequência, as classificadas como Fator de Impacto Médio 37,94% segundo conjunto de cinco palavras com maior frequência e as classificadas como Fator de Impacto Baixo 25,61% terceiro conjunto de cinco palavras com maior frequência.

5.3.7 Número de mensagens em meses por Fator de Impacto do Corpus Hotel Urbano

Na figura 55 é apresentada a forma que foi contabilizada as mensagens mais frequentes por meses presentes nas notícias publicadas na página de fãs Hotel Urbano classificado como fator de impacto alto, médio e baixo.

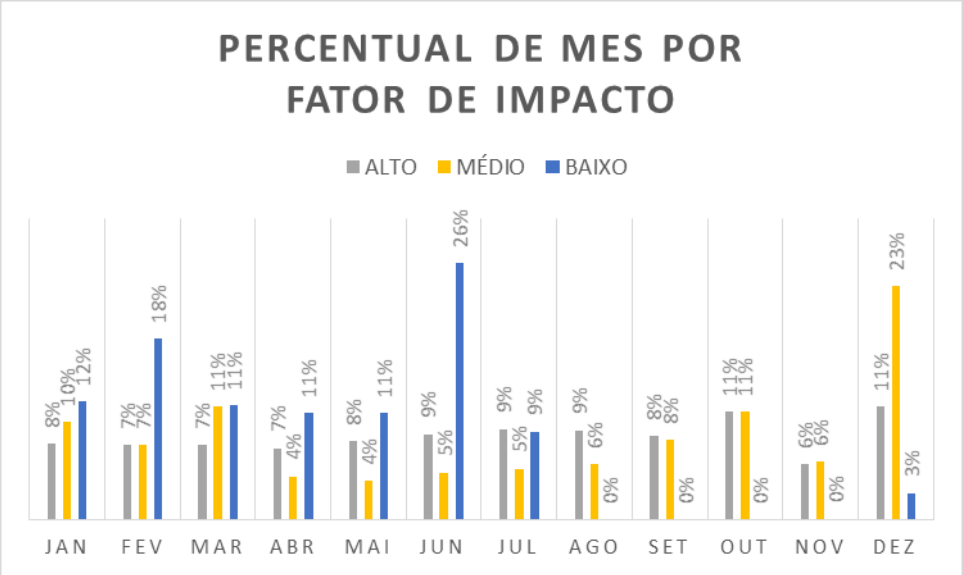


Figura 55 - Percentual de notícias em meses por Fator de Impacto do Corpus *Hotel Urbano*

Conforme os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por meses presente na página de fãs do Hotel Urbano, revela que em notícias classificadas como fator de impacto alto os meses com mais mensagens foram janeiro, julho e outubro. No entanto em notícias classificadas cm fator de impacto médio e baixo é evidenciado o mês de fevereiro.

5.3.8 Número de mensagens em cada dia da semana por Fator de Impacto do Corpus Hotel Urbano

Na figura 56 é apresentada a forma que foi contabilizada a frequência de mensagens por dia da semana presentes nas notícias publicadas na página de fãs Hotel Urbano classificado como fator de impacto alto, médio e baixo.

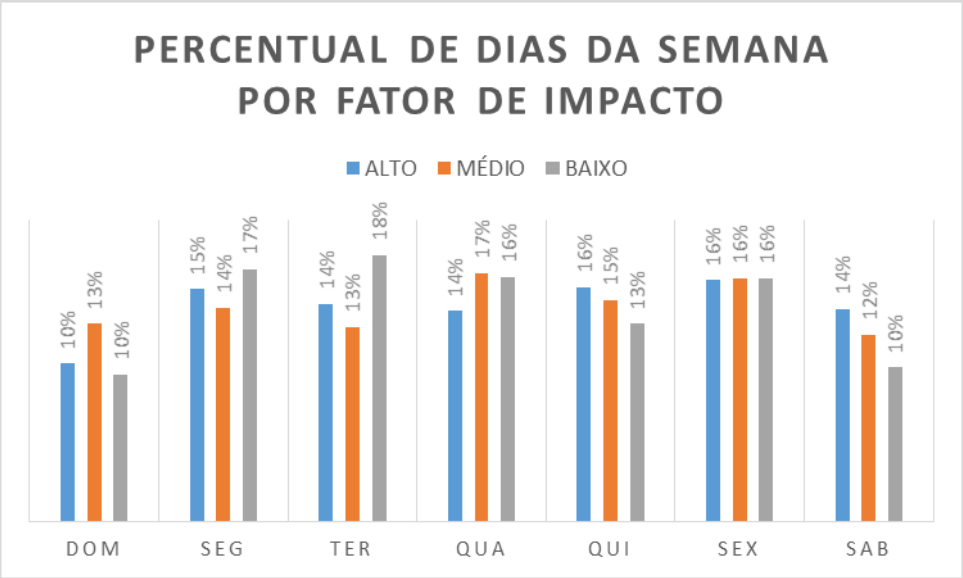


Figura 56 - Percentual de notícias por semana por Fator de Impacto do Corpus *Hotel Urbano*

Através dos resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por dia da semana presente na página de fãs do Hotel Urbano, revela que em notícias classificadas como fator de impacto alto o dia da semana vencedor foi sexta-feira. Contudo em notícias classificadas como fator de impacto médio e baixo o dia da semana foi terça-feira e quarta-feira.

5.3.9 Número de mensagens em turnos por Fator de Impacto do Corpus Hotel Urbano

Na figura 57 é apresentada a forma que foi contabilizada a frequência de mensagens por turno presentes nas notícias publicadas na página de fãs Hotel Urbano classificado como fator de impacto alto, médio e baixo.

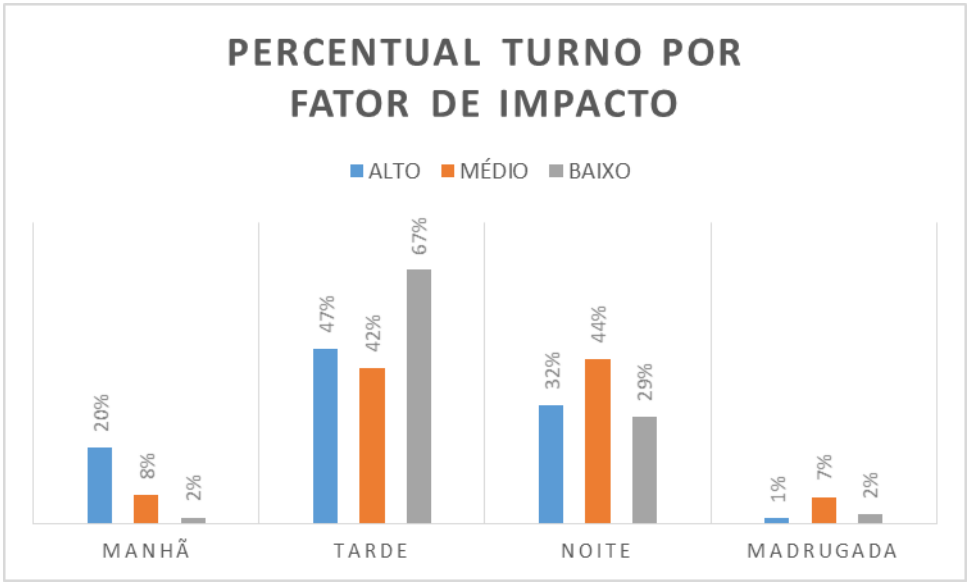


Figura 57 – Percentual de notícias por turno e por Fator de Impacto do Corpus *Hotel Urbano*

Segundo os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por turno presente na página de fãs do Hotel Urbano, revela que tanto em notícias classificadas como fator de impacto alto, médio e baixo o turno da tarde concentra uma maior quantidade de notícias, segundo também do turno da noite.

5.3.10 Número de tipos de mensagens por Fator de Impacto do Corpus Hotel Urbano

Na figura 58 é apresentada a forma que foi contabilizada a frequência de tipos de mensagens presentes nas notícias publicadas na página de fãs Hotel Urbano classificado como fator de impacto alto, médio e baixo.

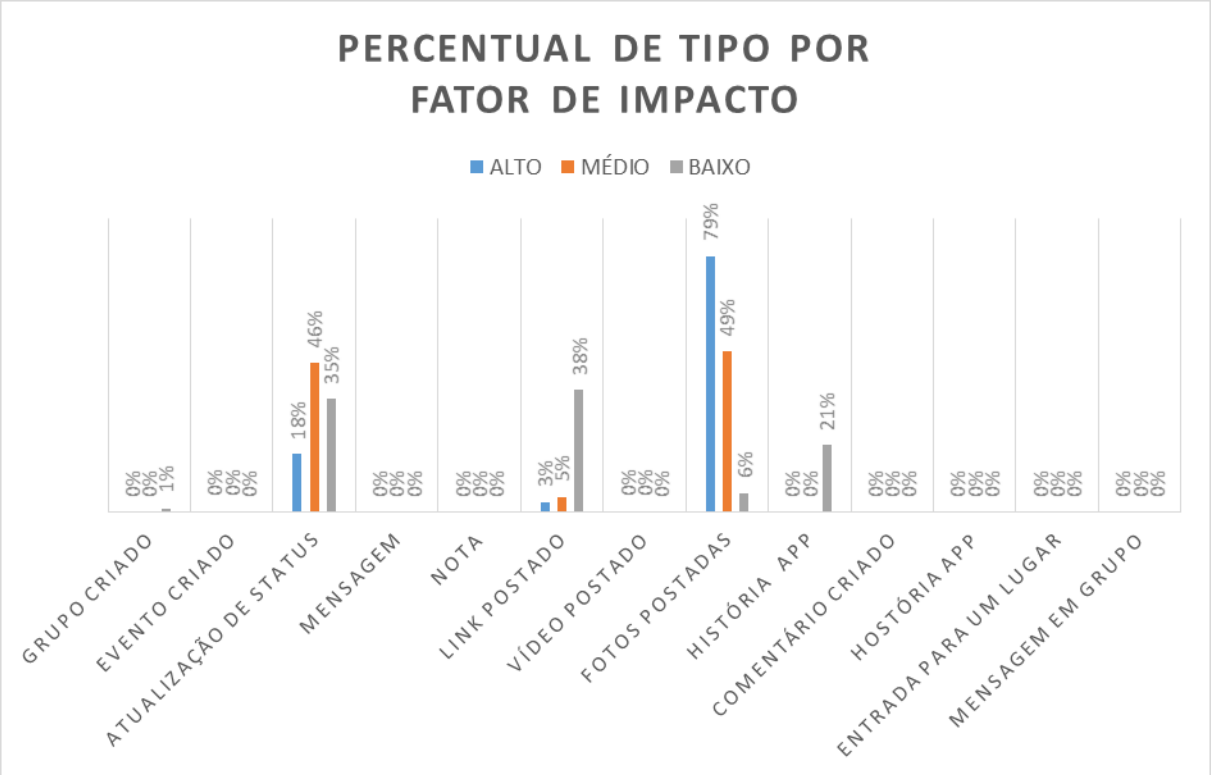


Figura 58 - Percentual de notícias por tipo e por Fator de Impacto do Corpus *Hotel Urbano*

Consoante os resultados apresentados acima pode-se perceber que se tratando de notícias classificadas como fator de impacto alto e médio o tipo de notícia que mais ocorre é de Atualização de Satus. E em notícias classificadas como fator de impacto baixo o tipo evidenciado foi o de link postado.

5.4 Atributos que Exercem Influência no Fator de Impacto do corpus Garoto.

Da mesma forma, nesta tarefa foram analisados os seguintes elementos com o objetivo de identificar os fatores que exercem influência no fator de impacto proposto no presente trabalho, dentre eles são listados os seguintes:

1. Número de palavras por Fator de Impacto;
2. Número de curtidas, comentários e compartilhamento por Fator de Impacto;
3. Número de Unigram por Fator de Impacto;
4. Número de Bigram por Fator de Impacto;
5. Número de Trigram por Fator de Impacto;
6. Número de Ngram por Fator de Impacto;
7. Número de mensagens em cada mês por Fator de Impacto;
8. Número de mensagens em cada dia da semana por Fator de Impacto;
9. Número de mensagens em cada turno por Fator de Impacto;
10. Número de tipos de mensagens por Fator de Impacto.

5.4.1 Número de palavras por Fator de Impacto do Corpus Garoto

Na figura 59 é apresentada a contabilização total de palavras presentes nas notícias publicadas na página de fãs do Garoto classificadas como fator de impacto alto, médio e baixo.

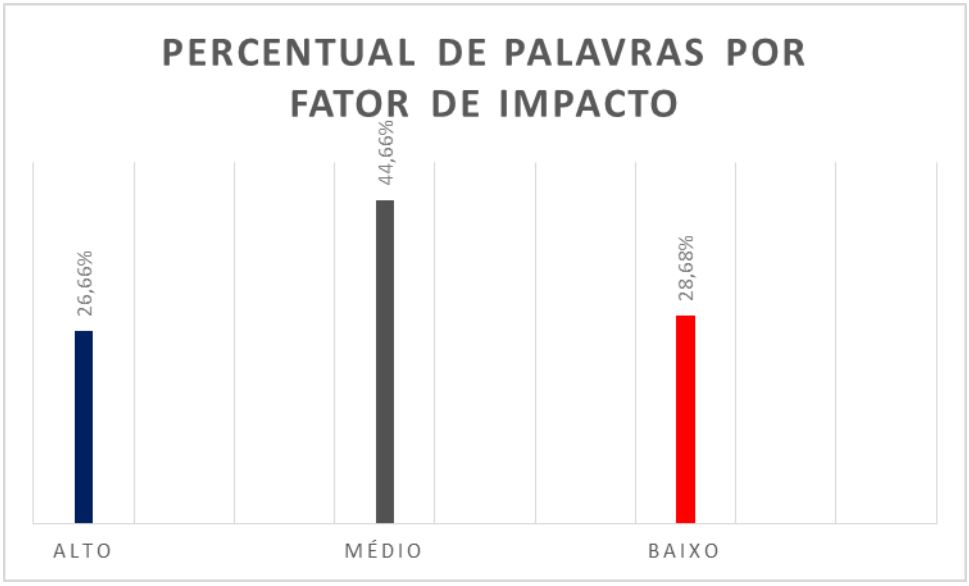


Figura 59 - Percentual de Palavras por Fator de Impacto do Corpus *Garoto*

De acordo com os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 26,66% de número de palavras, 44,66% do número de palavras classificadas como Fator de Impacto Médio e 28,68% do número de palavras classificadas como Fator de Impacto Baixo.

5.4.2 Número de curtidas, comentários e compartilhamento por Fator de Impacto do Corpus Garoto

Na figura 60 é apresentada a forma que foi contabilizada as seguintes interações: número de curtidas, comentários e compartilhamentos presentes nas notícias publicadas na página de fãs Garoto classificadas como fator de impacto alto, médio e baixo.

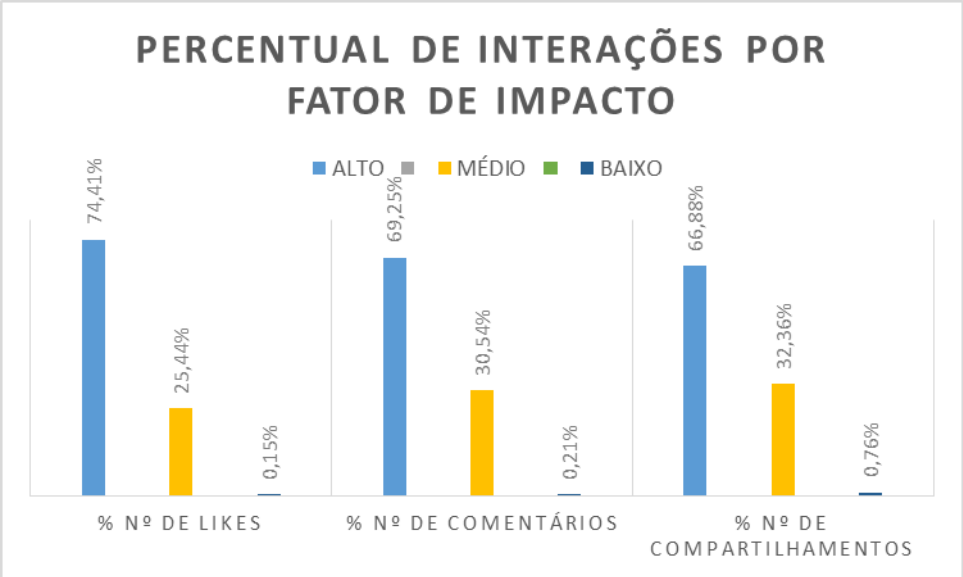


Figura 60 – Percentual de Interações por Fator de Impacto do Corpus *Garoto*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 74,41% do número de likes, 69,25% do número de comentários e 66,88% do número de compartilhamentos.

De outra forma significa que no corpus de notícias extraídas da página de fãs Garoto quanto maior o número de likes, comentários e compartilhamentos presentes nas notícias divulgadas na página de fãs da rede social Facebook maior será seu Fator de Impacto.

5.4.3 Número de Unigram por Fator de Impacto do Corpus Garoto

Na sequência é apresentada a forma que foi contabilizada as palavras mais frequentes presentes nas notícias publicadas na página de fãs Garoto classificadas como fator de impacto alto, médio e baixo.

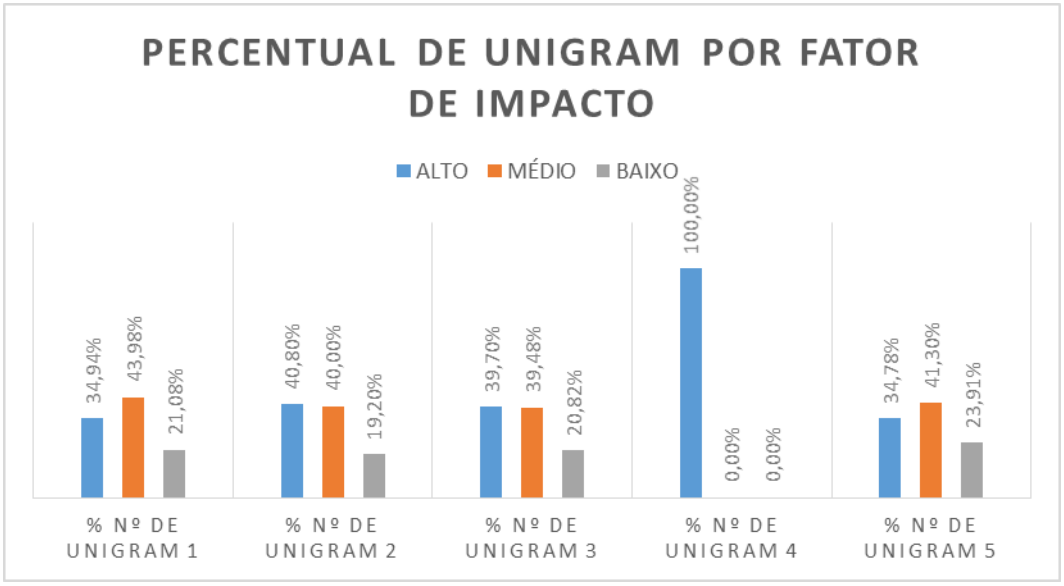


Figura 61 - Percentual de Unigram por Fator de Impacto do Corpus *Garoto*

Conforme os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 34,94% primeira palavra com maior frequência, 40,80% segunda palavra com maior frequência, 39,70% terceira palavra com maior frequência, 100% quarta e 34,78% quinta palavra com maior frequência.

De outra forma significa que no corpus de notícias extraídas da página de fãs da Garoto as palavras com maior frequência ocorrem de forma similar as notícias classificadas como fator de impacto Médio e somente na quarta palavra com maior frequência houve a ocorrência com 100%.

5.4.4 Número de Bigram por Fator de Impacto do Corpus Garoto

Na figura 62 é apresentada a forma que foi contabilizada os pares de palavras mais frequentes presentes nas notícias publicadas na página de fãs Garoto classificadas como fator de impacto alto, médio e baixo.

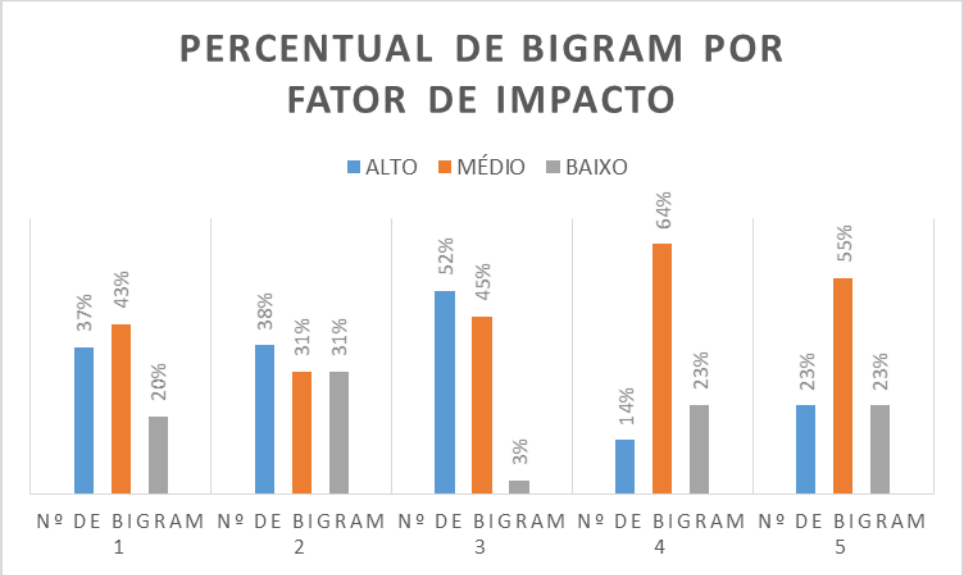


Figura 62 - Percentual de Bigram por Fator de Impacto do Corpus *Garoto*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 37% primeiro conjunto de duas palavras com maior frequência, 51,61% segundo conjunto de duas com maior frequência, 37,78% terceiro conjunto de duas com maior frequência, 33,71% quarto conjunto de duas palavras com maior frequência e 22,50% quinta conjunto de duas palavras com maior frequência.

5.4.5 Número de Trigram por Fator de Impacto do Corpus Garoto

Na figura 63 é apresentada a forma que foi contabilizada os conjuntos de três palavras mais frequentes presentes nas notícias publicadas na página de fãs Garoto classificadas como fator de impacto alto, médio e baixo.

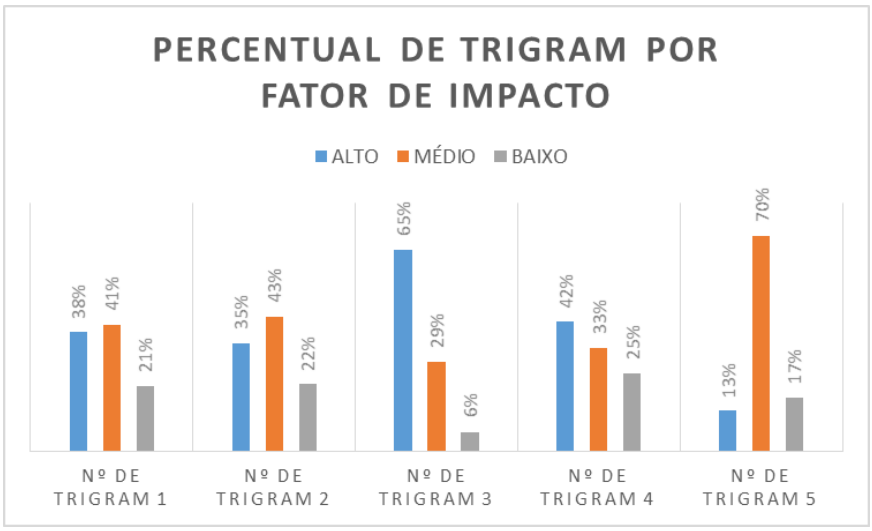


Figura 63 - Percentual de Trigram por Fator de Impacto do Corpus *Garoto*

Segundo os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 41,67% primeiro conjunto de três palavras com maior frequência, 13,04% segundo conjunto de três com maior frequência, 65,15% terceiro conjunto de três palavras com maior frequência, 34,93% quarto conjunto de três palavras com maior frequência e 38,30% quinto conjunto de três palavras com maior frequência.

5.4.6 Número de Ngram por Fator de Impacto do Corpus Garoto

Na figura 64 é apresentada a forma que foi contabilizada os conjuntos de cinco e dez palavras mais frequentes presentes nas notícias publicadas na página de fãs Garoto classificadas como fator de impacto alto, médio e baixo.

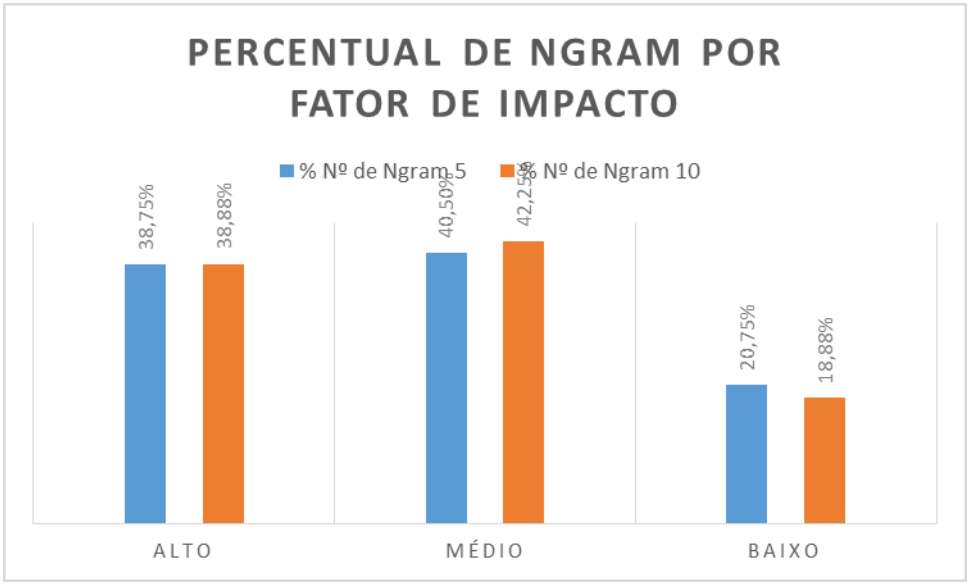


Figura 64 - Percentual de Ngram por Fator de Impacto do Corpus *Garoto*

Consoante os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 38,75% primeiro conjunto de cinco palavras com maior frequência, as classificadas como Fator de Impacto Médio 40,50% segundo conjunto de cinco palavras com maior frequência e as classificadas como Fator de Impacto Baixo 20,75% terceiro conjunto de cinco palavras com maior frequência.

5.4.7 Número de mensagens em meses por Fator de Impacto do Corpus Garoto

Na figura 65 é apresentada a forma que foi contabilizada as mensagens mais frequentes por meses presentes nas notícias publicadas na página de fãs Garoto classificadas como fator de impacto alto, médio e baixo.

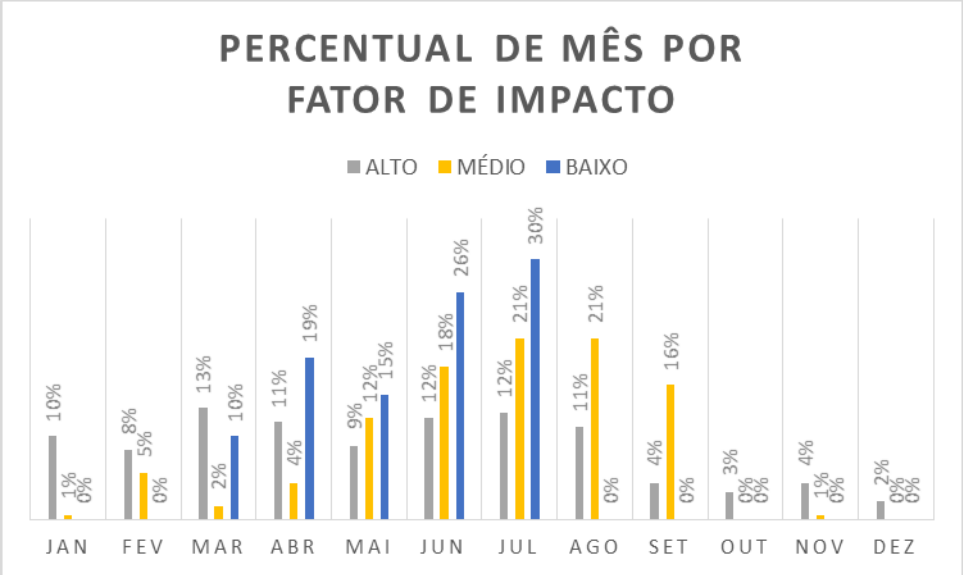


Figura 65 - Percentual de Notícias em meses e por Fator de Impacto do Corpus *Garoto*

De acordo com os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por meses presente na página de fãs do Garoto, revela que em notícias classificadas como fator de impacto alto os meses com mais mensagens foram janeiro, julho e outubro. No entanto em notícias classificadas cm fator de impacto médio e baixo é evidenciado o mês de fevereiro.

5.4.8 Número de mensagens em cada dia da semana por Fator de Impacto do Corpus Garoto

Na figura 66 é apresentada a forma que foi contabilizada a frequência de mensagens por dia da semana presentes nas notícias publicadas na página de fãs Garoto classificadas como fator de impacto alto, médio e baixo.

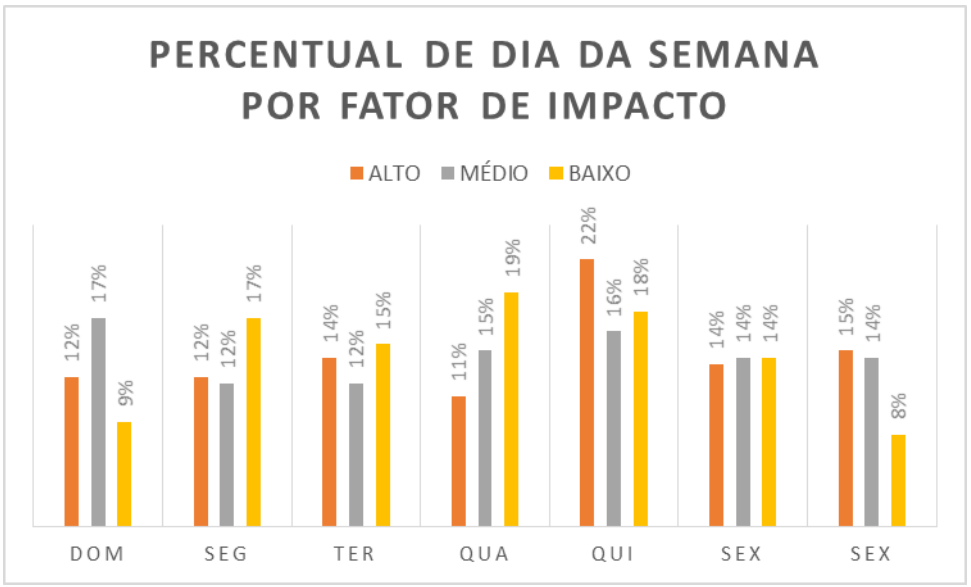


Figura 66 - Percentual de Notícias por dia da semana e por Fator de Impacto do Corpus *Garoto*

Através dos resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por dia da semana presente na página de fãs do Garoto, revela que em notícias classificadas como fator de impacto alto o dia da semana vencedor foi quinta-feira. Contudo em notícias classificadas como fator de impacto médio e baixo o dia da semana foi domingo e quarta-feira.

5.4.9 Número de mensagens em turnos por Fator de Impacto do Corpus Garoto

Na figura 67 é apresentada a forma que foi contabilizada a frequência de mensagens por turno presentes nas notícias publicadas na página de fãs Garoto classificadas como fator de impacto alto, médio e baixo.

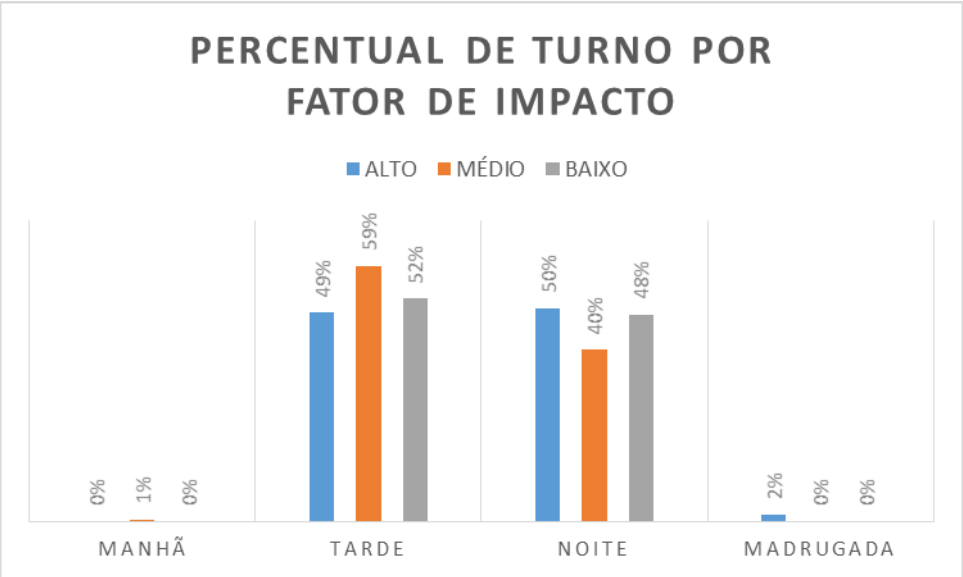


Figura 67 - Percentual de Notícias por turno e por Fator de Impacto do Corpus *Garoto*

Conforme os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por turno presente na página de fãs do Garoto, revela que tanto em notícias classificadas como fator de impacto alto, médio e baixo o turno da tarde concentra uma maior quantidade de notícias, segundo também do turno da noite.

5.4.10 Número de tipos de mensagens por Fator de Impacto do Corpus Garoto

Na figura 68 é apresentada a forma que foi contabilizada a frequência de tipos de mensagens presentes nas notícias publicadas na página de Garoto urbano classificadas como fator de impacto alto, médio e baixo.

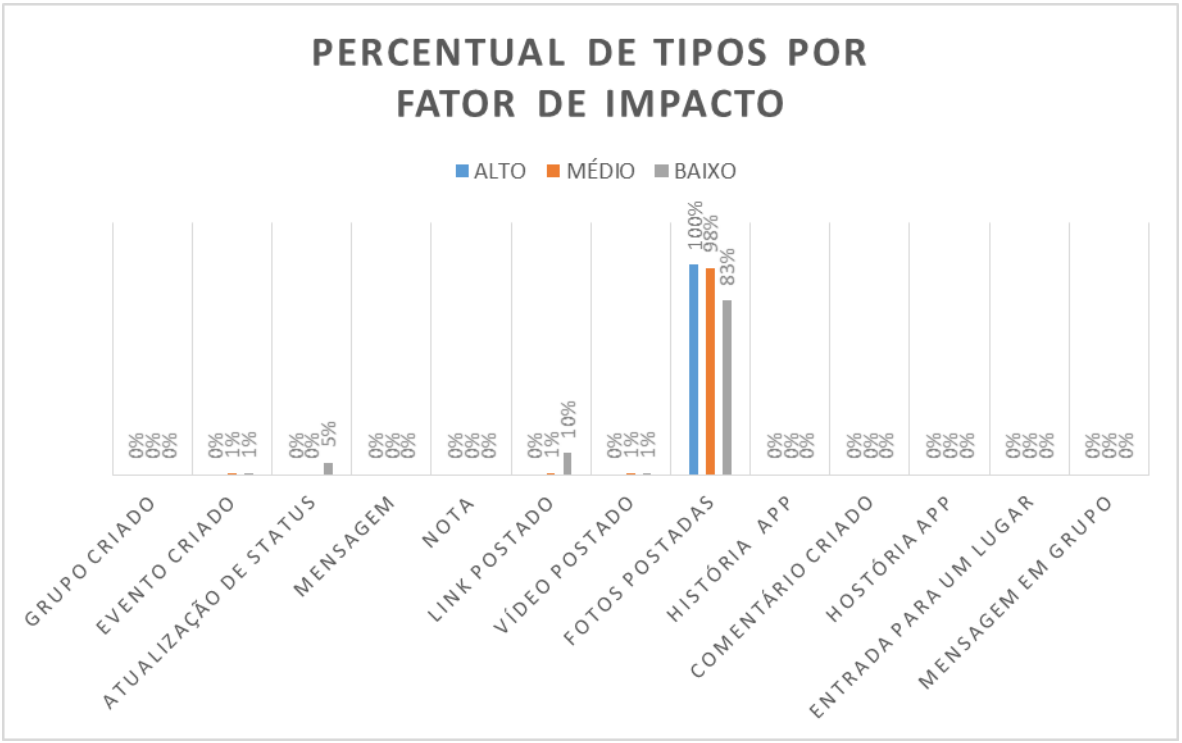


Figura 68 - Percentual de Notícias por tipo e por Fator de Impacto do Corpus *Garoto*

Através dos resultados apresentados acima pode-se perceber que se tratando de notícias classificadas como fator de impacto alto e médio o tipo de notícia que mais ocorre é de Atualização de Satus. E em notícias classificadas como fator de impacto baixo o tipo evidenciado foi o de link postado.

5.5 Atributos que exercem influência no fator de impacto do corpus Lacta.

Da mesma forma apresentada acima, nesta tarefa foram analisados os seguintes elementos com o objetivo de identificar os fatores que exercem influência no fator de impacto proposto no presente trabalho, dentre eles são listados os seguintes:

1. Número de palavras por Fator de Impacto;
2. Número de curtidas, comentários e compartilhamento por Fator de Impacto;
3. Número de Unigram por Fator de Impacto;
4. Número de Bigram por Fator de Impacto;
5. Número de Trigram por Fator de Impacto;
6. Número de Ngram por Fator de Impacto;
7. Número de mensagens em cada mês por Fator de Impacto;
8. Número de mensagens em cada dia da semana por Fator de Impacto;
9. Número de mensagens em cada turno por Fator de Impacto;
10. Número de tipos de mensagens por Fator de Impacto.

5.5.1 Número de palavras por Fator de Impacto do Corpus Lacta

Na figura 69 é apresentada a contabilização total de palavras presentes nas notícias publicadas na página de fãs do Lacta classificadas como fator de impacto alto, médio e baixo.

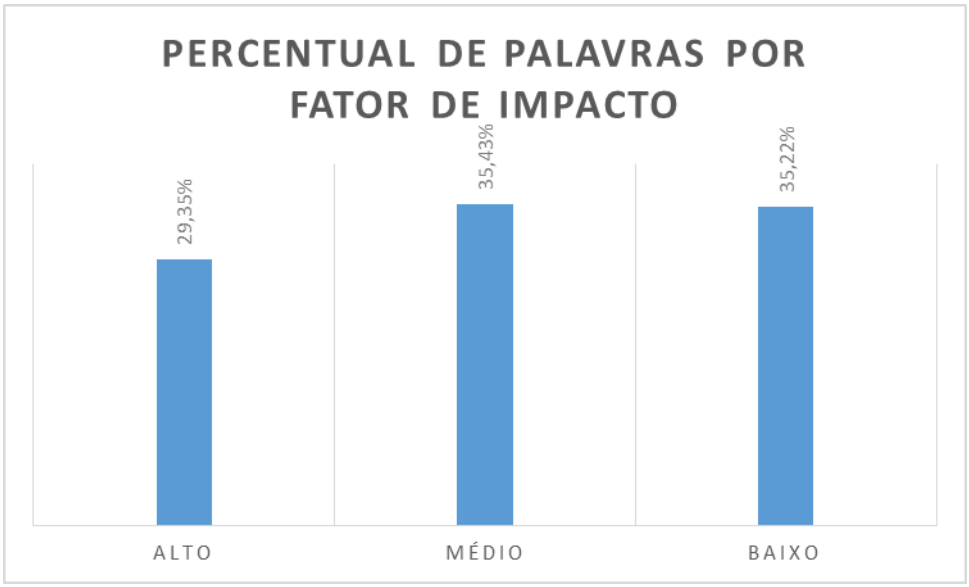


Figura 69 - Percentual de Palavras por Fator de Impacto do Corpus *Lacta*

Segundo os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 29,35% de número de palavras, 35,43% do número de palavras classificadas como Fator de Impacto Médio e 35,22% do número de palavras classificadas como Fator de Impacto Baixo.

5.5.2 Número de curtidas, comentários e compartilhamento por Fator de Impacto do Corpus Lacta

Na figura 70 é apresentada a forma que foi contabilizada as seguintes interações: número de curtidas, comentários e compartilhamentos presentes nas notícias publicadas na página de fãs Lacta classificadas como fator de impacto alto, médio e baixo.

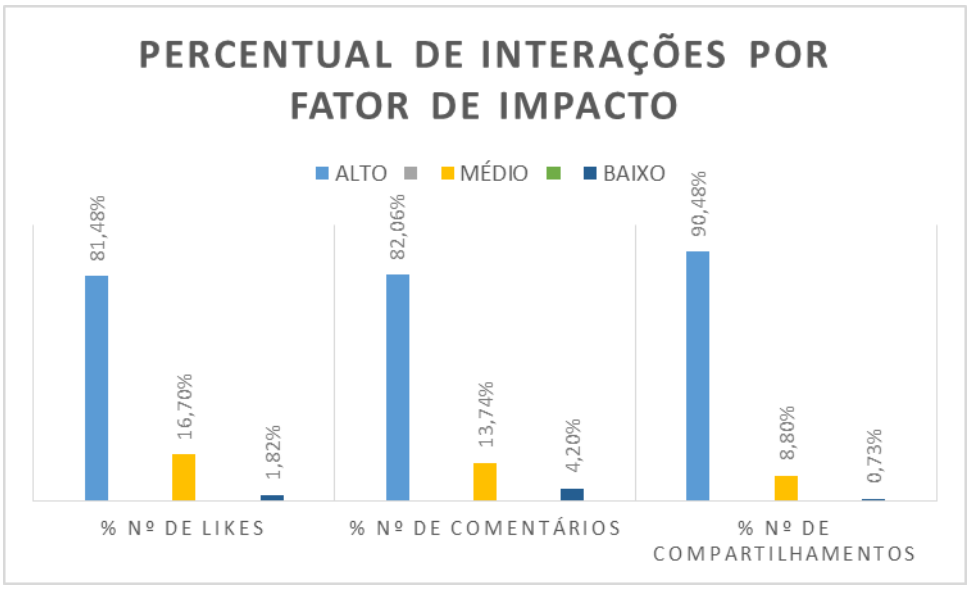


Figura 70 – Percentual de Interações por Fator de Impacto do Corpus *Lacta*

Consoante os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 81,48% do número de likes, 82,06% do número de comentários e 90,48% do número de compartilhamentos.

De outra forma significa que no corpus de notícias extraídas da página de fãs da Lacta quanto maior o número de likes, comentários e compartilhamentos presentes nas notícias divulgadas na página de fãs da rede social Facebook maior será seu Fator de Impacto.

Observa-se também que o número de compartilhamentos supera o percentual de likes e comentários em tais notícias.

5.5.3 Número de Unigram por Fator de Impacto do Corpus Lacta

Na figura 71 é apresentada a forma que foi contabilizada as palavras mais frequentes presentes nas notícias publicadas na página de fãs Lacta classificadas como fator de impacto alto, médio e baixo.

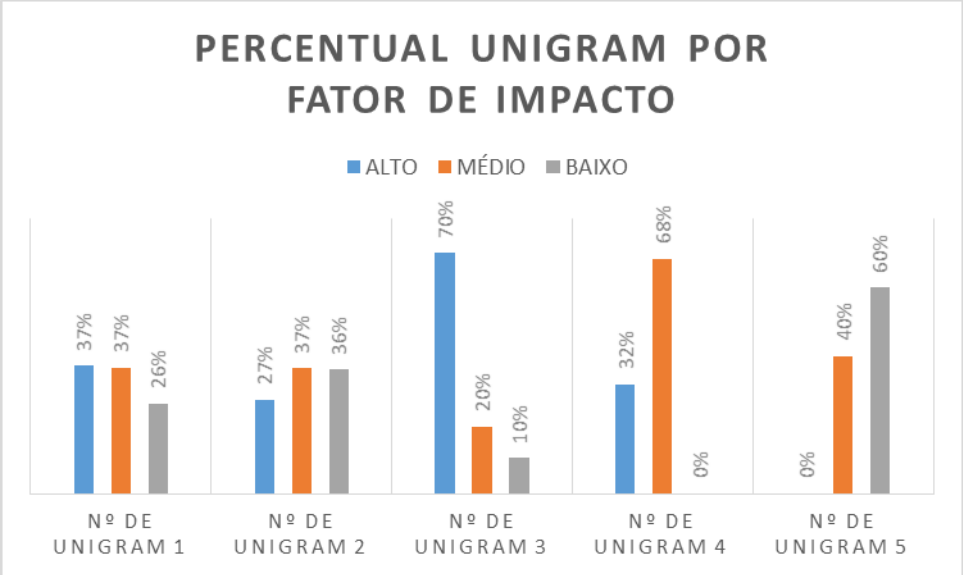


Figura 71 - Percentual de Unigram por Fator de Impacto do Corpus *Lacta*

De acordo com os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 37% primeira palavra com maior frequência, 27 % segunda palavra com maior frequência, 70% terceira palavra com maior frequência, 32% quarta palavra com maior frequência e 0% quinta palavra com maior frequência.

5.5.4 Número de Bigram por Fator de Impacto do Corpus Lacta

Na figura 72 é apresentada a forma que foi contabilizada os pares de palavras mais frequentes presentes nas notícias publicadas na página de fãs Lacta classificadas como fator de impacto alto, médio e baixo.

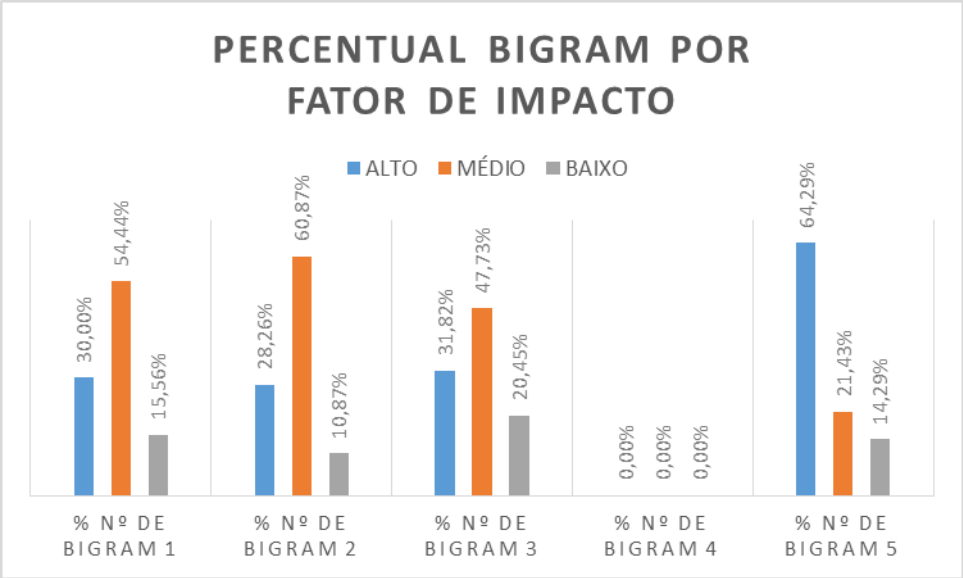


Figura 72 - Percentual de Bigram por Fator de Impacto do Corpus *Lacta*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 30,00% primeiro conjunto de duas palavras com maior frequência, 28,26% segundo conjunto de duas com maior frequência, 31,82% terceiro conjunto de duas com maior frequência, 0,00% quarto conjunto de duas palavras com maior frequência e 64,29% quinta conjunto de duas palavras com maior frequência.

5.5.5 Número de Trigram por Fator de Impacto

Na figura 73 é apresentada a forma que foi contabilizada os conjuntos de três palavras mais frequentes presentes nas notícias publicadas na página de fãs Lacta classificadas como fator de impacto alto, médio e baixo.

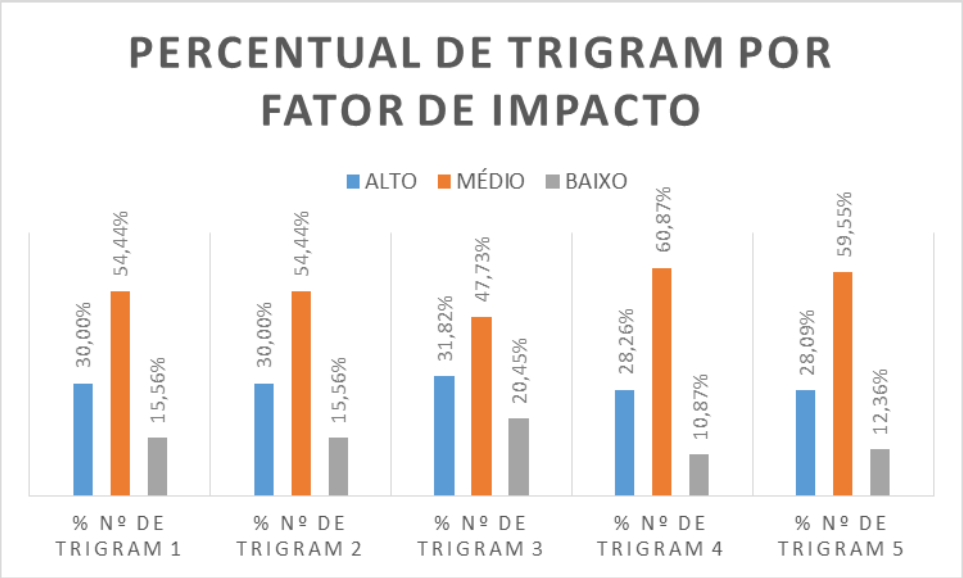


Figura 73 - Percentual de Trigram por Fator de Impacto do Corpus *Lacta*

Conforme os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 30,00% primeiro conjunto de três palavras com maior frequência, 30,00% segundo conjunto de três com maior frequência, 31,82% terceiro conjunto de três palavras com maior frequência, 28,26% quarto conjunto de três palavras com maior frequência e 28,09% quinto conjunto de três palavras com maior frequência.

5.5.6 Número de Ngram por Fator de Impacto

Na figura 74 é apresentada a forma que foi contabilizada os conjuntos de cinco e dez palavras mais frequentes presentes nas notícias publicadas na página de fãs Lacta classificadas como fator de impacto alto, médio e baixo.

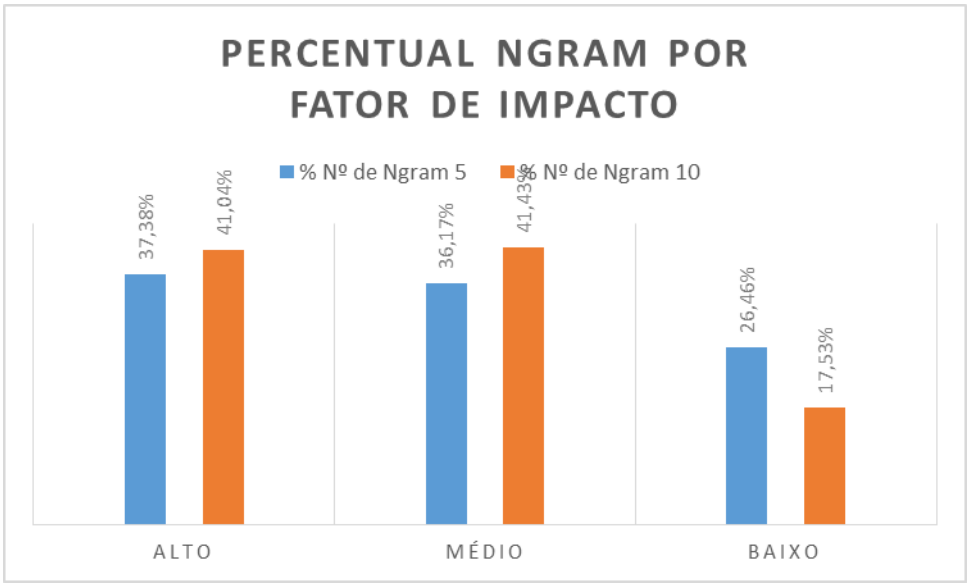


Figura 74 - Percentual de Ngram por Fator de Impacto do Corpus *Lacta*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 37,38% primeiro conjunto de cinco palavras com maior frequência, as classificadas como Fator de Impacto Médio 36,41% segundo conjunto de cinco palavras com maior frequência e as classificadas como Fator de Impacto Baixo 26,46% terceiro conjunto de cinco palavras com maior frequência.

5.5.7 Número de mensagens em meses por Fator de Impacto

Na figura 75 é apresentada a forma que foi contabilizada as mensagens mais frequentes por meses presentes nas notícias publicadas na página de fãs Lacta classificadas como fator de impacto alto, médio e baixo.

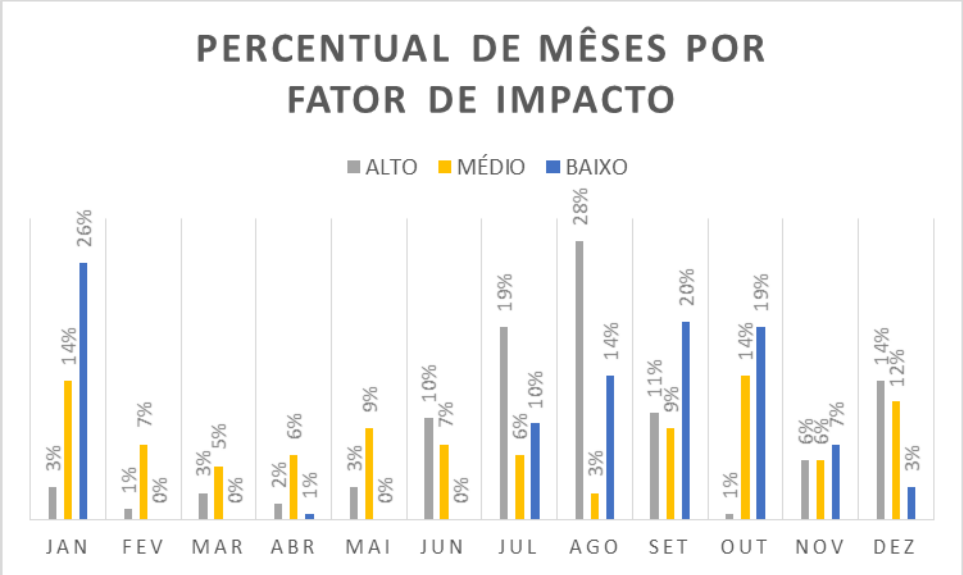


Figura 75 - Percentual de Notícias em meses por Fator de Impacto do Corpus *Lacta*

Segundo os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por meses presente na página de fãs do Lacta, revela que em notícias classificadas como fator de impacto alto os meses com mais mensagens foram janeiro, julho e outubro. No entanto em notícias classificadas cm fator de impacto médio e baixo é evidenciado o mês de fevereiro.

5.5.8 Número de mensagens em cada dia da semana por Fator de Impacto

Na figura 76 é apresentada a forma que foi contabilizada a frequência de mensagens por dia da semana presentes nas notícias publicadas na página de fãs Lacta classificadas como fator de impacto alto, médio e baixo.

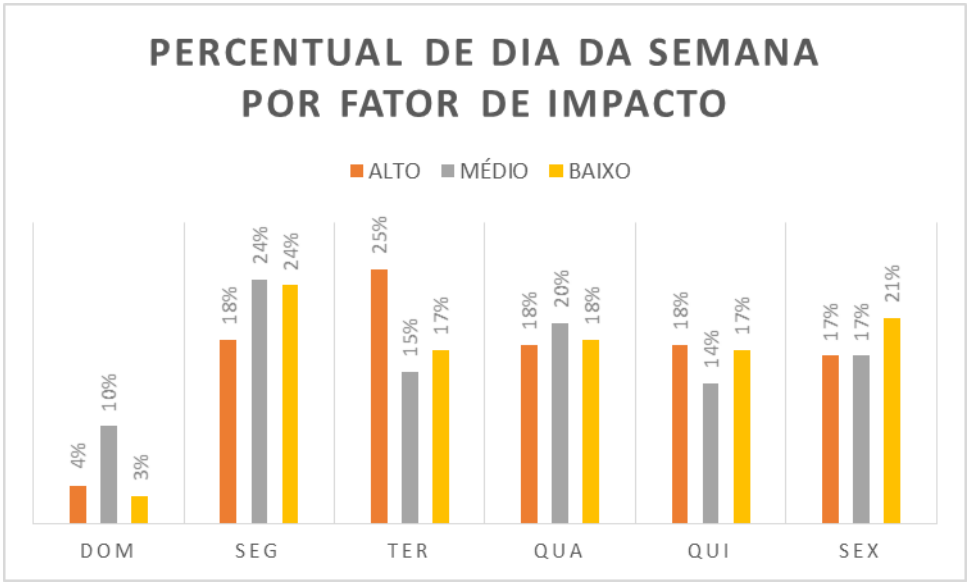


Figura 76 - Percentual de Notícias por dia da semana e por Fator de Impacto do Corpus *Lacta*

Consoante os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por dia da semana presente na página de fãs do Lacta, revela que em notícias classificadas como fator de impacto alto o dia da semana vencedor foi segunda. Contudo em notícias classificadas como fator de impacto médio e baixo o dia da semana foi quarta.

5.5.9 Número de mensagens em turnos por Fator de Impacto

Na figura 77 é apresentada a forma que foi contabilizada a frequência de mensagens por turno presentes nas notícias publicadas na página de fãs Lacta classificadas como fator de impacto alto, médio e baixo.

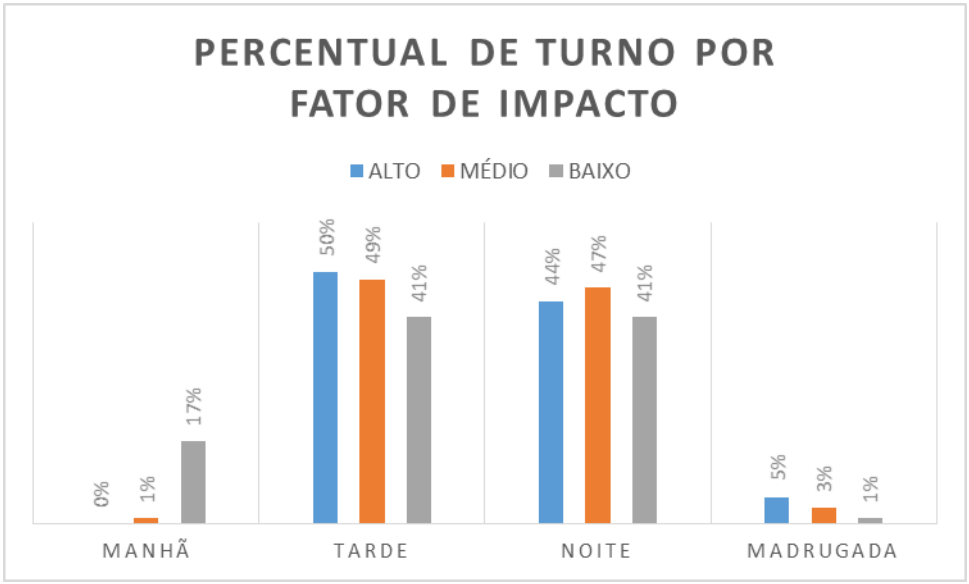


Figura 77 - Percentual de Notícias por turno e por Fator de Impacto do Corpus *Lacta*

De acordo com os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por turno presente na página de fãs do Lacta, revela que tanto em notícias classificadas como fator de impacto alto, médio e baixo o turno da tarde concentra uma maior quantidade de notícias, segundo também do turno da noite.

5.5.10 Número de tipos de mensagens por Fator de Impacto

Na figura 78 é apresentada a forma que foi contabilizada a frequência de tipos de mensagens presentes nas notícias publicadas na página de fãs Lacta classificadas como fator de impacto alto, médio e baixo.

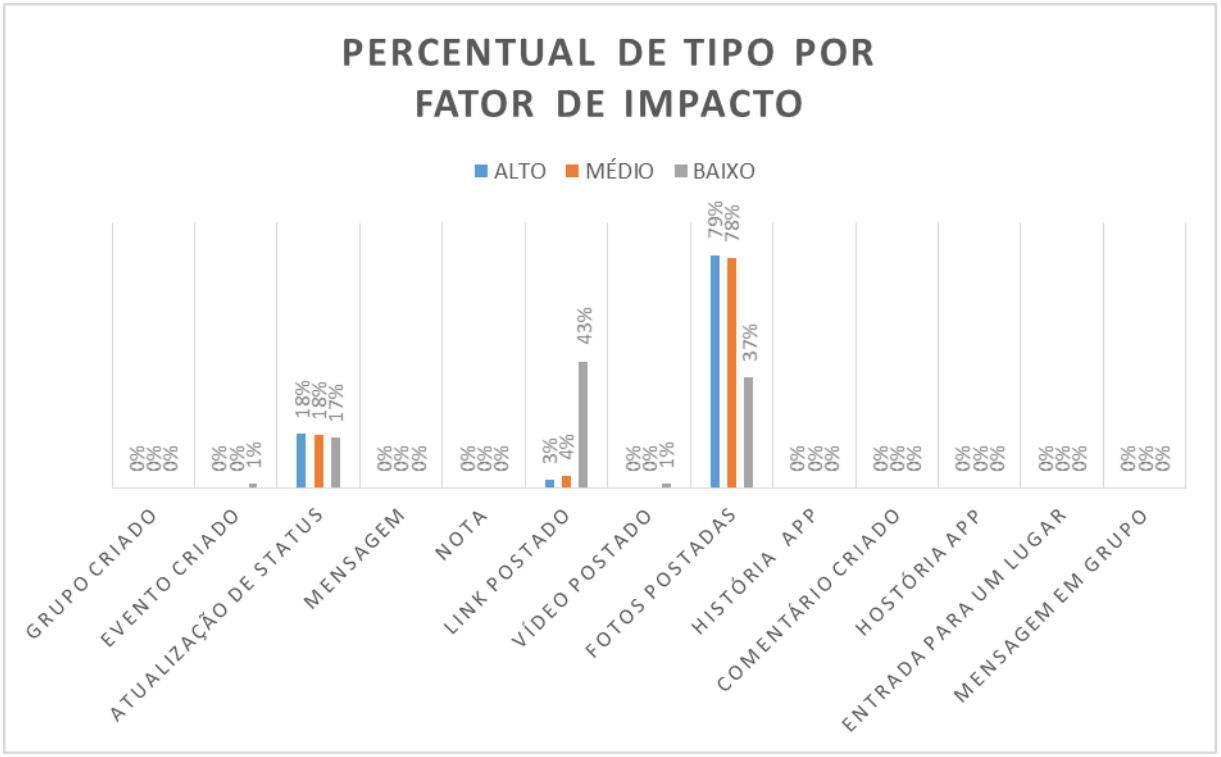


Figura 78 - Percentual de Notícias por tipo e por Fator de Impacto do Corpus *Lacta*

Através dos resultados apresentados acima pode-se perceber que se tratando de notícias classificadas como fator de impacto alto e médio o tipo de notícia que mais ocorre é de Atualização de Satus. E em notícias classificadas como fator de impacto baixo o tipo evidenciado foi o de link postado.

5.6 Atributos que exercem influência no fator de impacto do corpus União Corpora.

Da mesma forma, nesta tarefa foram analisados os seguintes elementos com o objetivo de identificar os fatores que exercem influência no fator de impacto proposto no presente trabalho, dentre eles são listados os seguintes:

1. Número de palavras por Fator de Impacto;
2. Número de curtidas, comentários e compartilhamento por Fator de Impacto;
3. Número de Unigram por Fator de Impacto;
4. Número de Bigram por Fator de Impacto;
5. Número de Trigram por Fator de Impacto;
6. Número de Ngram por Fator de Impacto;
7. Número de mensagens em cada mês por Fator de Impacto;
8. Número de mensagens em cada dia da semana por Fator de Impacto;
9. Número de mensagens em cada turno por Fator de Impacto;
10. Número de tipos de mensagens por Fator de Impacto.

5.6.1 Número de palavras por Fator de Impacto do Corpus União Corpora

Na figura 79 é apresentada a contabilização total de palavras presentes nas notícias publicadas na página de fãs da Coca Cola classificada como fator de impacto alto, médio e baixo.

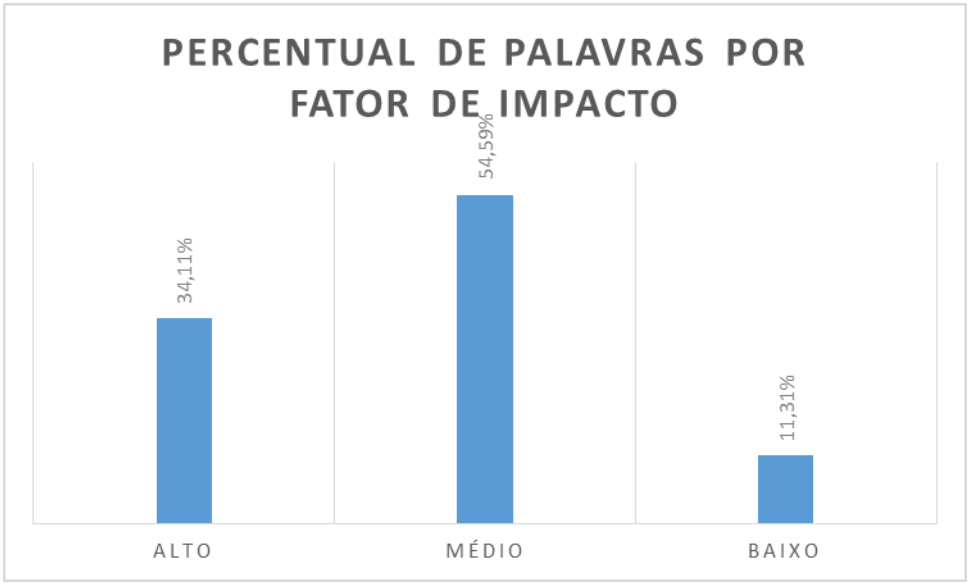


Figura 79 - Percentual de Palavras por Fator de Impacto do Corpus *União Corpora*

Conforme os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 34,11% de número de palavras, 54,59% do número de palavras classificadas como Fator de Impacto Médio e 11,31% do número de palavras classificadas como Fator de Impacto Baixo.

5.6.2 Número de curtidas, comentários e compartilhamento por Fator de Impacto do Corpus *União Corpora*.

Na figura 80 é apresentada a forma que foi contabilizada as seguintes interações: número de curtidas, comentários e compartilhamentos presentes nas notícias de todas as páginas de fãs classificado como fator de impacto alto, médio e baixo.

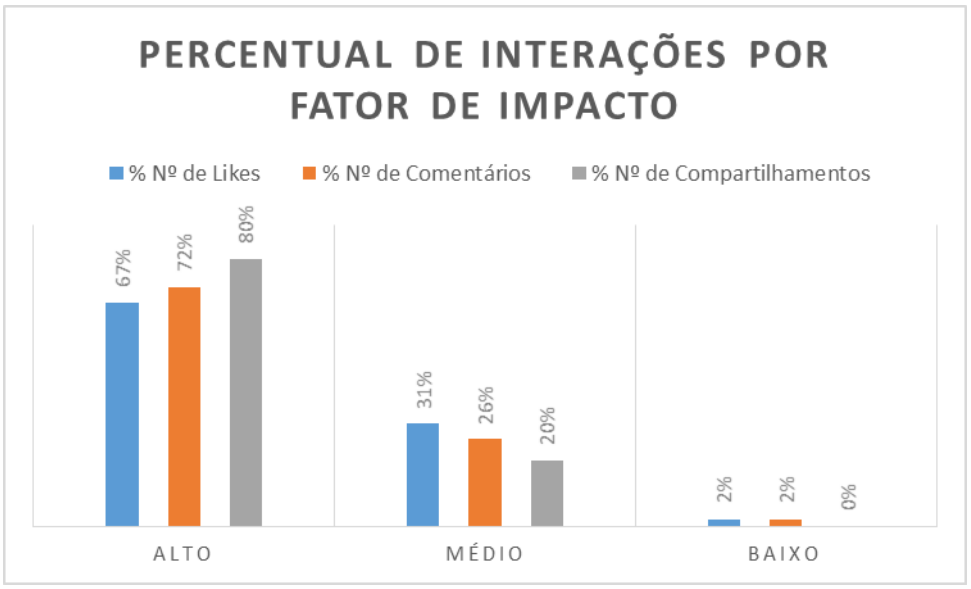


Figura 80 – Percentual de Interações por Fator de Impacto do Corpus *União Corpora*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 67% do número de likes, 72% do número de comentários e 80% do número de compartilhamentos.

De outra forma significa que no corpus de que unifica todas as notícias extraídas das páginas de fãs quanto maior o número de likes, comentários e compartilhamentos presentes nas notícias divulgadas na página de fãs da rede social Facebook maior será seu Fator de Impacto.

5.6.3 Número de Unigram por Fator de Impacto do Corpus *União Corpora*

Na figura 81 é apresentada a forma que foi contabilizada as palavras mais frequentes presentes nas notícias publicadas em todas as página de fãs classificadas como fator de impacto alto, médio e baixo.

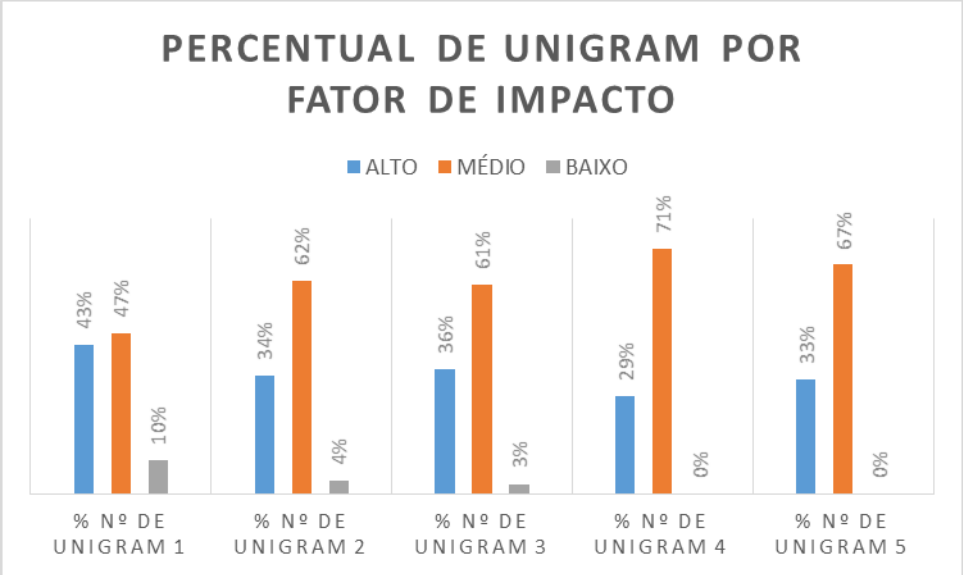


Figura 81 - Percentual de Unigram por Fator de Impacto do Corpus *União Corpora*

Segundo os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 43% de palavras mais frequentes, 34% com a segunda palavra mais frequente, 36% com a terceira palavra mais frequente, 29% com a quarta palavra mais frequente e 33% com a quinta palavra mais frequente.

É observado também que o maior resultado foi obtido nas classes com fator de impacto médio.

5.6.4 Número de Bigram por Fator de Impacto do Corpus *União Corpora*

Na figura 82 é apresentada a forma que foi contabilizada os pares de palavras mais frequentes presentes nas notícias publicadas em todas páginas de fãs classificados como fator de impacto alto, médio e baixo.

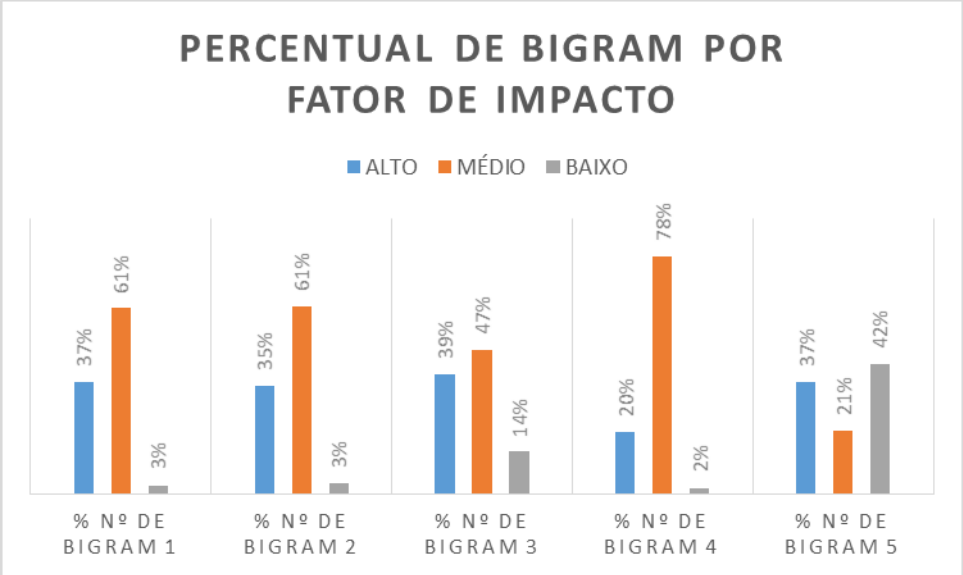


Figura 82 - Percentual de Bigram por Fator de Impacto do Corpus *União Corpora*

Consoante os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 37% primeiro conjunto de duas palavras com maior frequência, 35% segundo conjunto de duas com maior frequência, 39% terceiro conjunto de duas com maior frequência, 20% quarto conjunto de duas palavras com maior frequência e 37% quinta conjunto de duas palavras com maior frequência.

É observado também que o maior resultado foi obtido nas classes com fator de impacto médio.

5.6.5 Número de Trigram por Fator de Impacto do Corpus *União Corpora*

Na figura 83 é apresentada a forma que foi contabilizada os conjuntos de três palavras mais frequentes presentes nas notícias publicadas em todas páginas de fãs classificados como fator de impacto alto, médio e baixo.

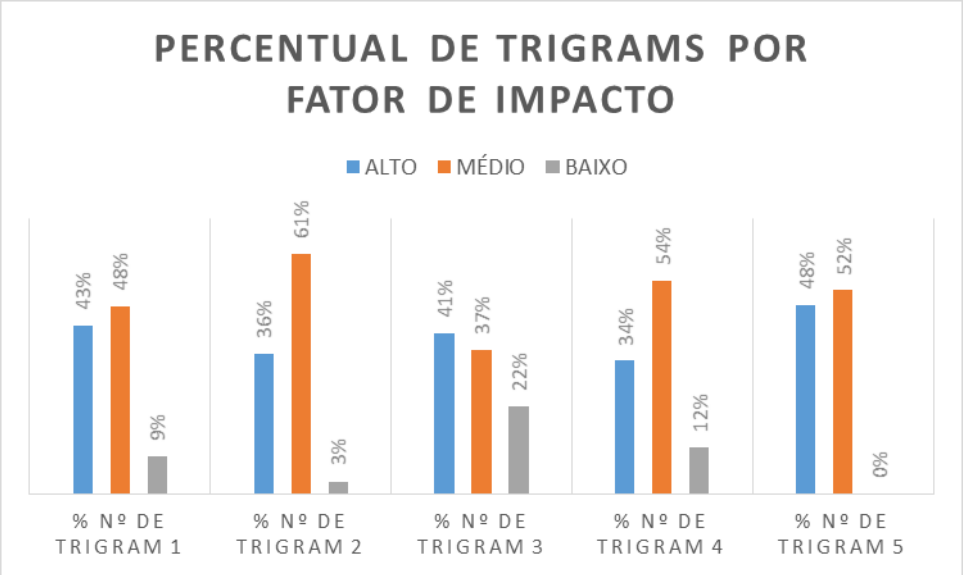


Figura 83 - Percentual de Trigram por Fator de Impacto do Corpus *União Corpora*

De acordo com os resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 43% primeiro conjunto de três palavras com maior frequência, 36% segundo conjunto, 41% terceiro conjunto, 34% quarto conjunto e 48% quinto conjunto de três palavras com maior frequência.

É observado também que o maior resultado foi obtido nas classes com fator de impacto médio.

5.6.6 Número de Ngram por Fator de Impacto do Corpus *União Corpora*

Na figura 84 é apresentada a forma que foi contabilizada os conjuntos de cinco e dez palavras mais frequentes presentes nas notícias publicadas em todas páginas de fãs classificados como fator de impacto alto, médio e baixo.

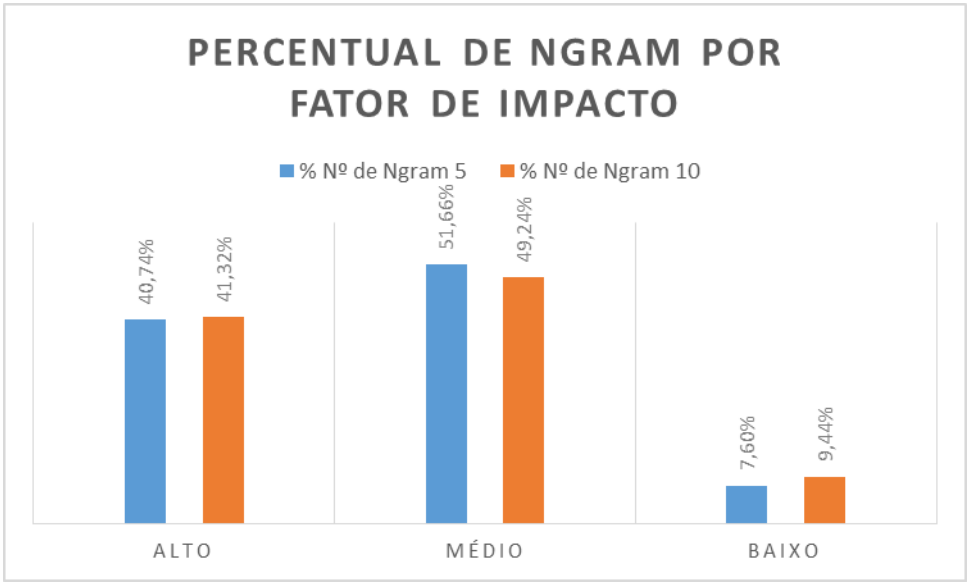


Figura 84 - Percentual de Ngram por Fator de Impacto do Corpus *União Corpora*

Através dos resultados apresentados pode-se perceber que notícias publicadas, classificadas como Fator de Impacto Alto, possuem 40,74% primeiro conjunto de cinco palavras com maior frequência e 41,32% utilizando o conjunto das dez palavras com maior frequência.

É observado também que o maior resultado foi obtido nas classes com fator de impacto médio.

5.6.7 Número de mensagens em meses por Fator de Impacto do Corpus Coca Cola

Na figura 85 é apresentada a forma que foi contabilizada as mensagens mais frequentes por meses presentes nas notícias publicadas em todas páginas de fãs classificadas como fator de impacto alto, médio e baixo.

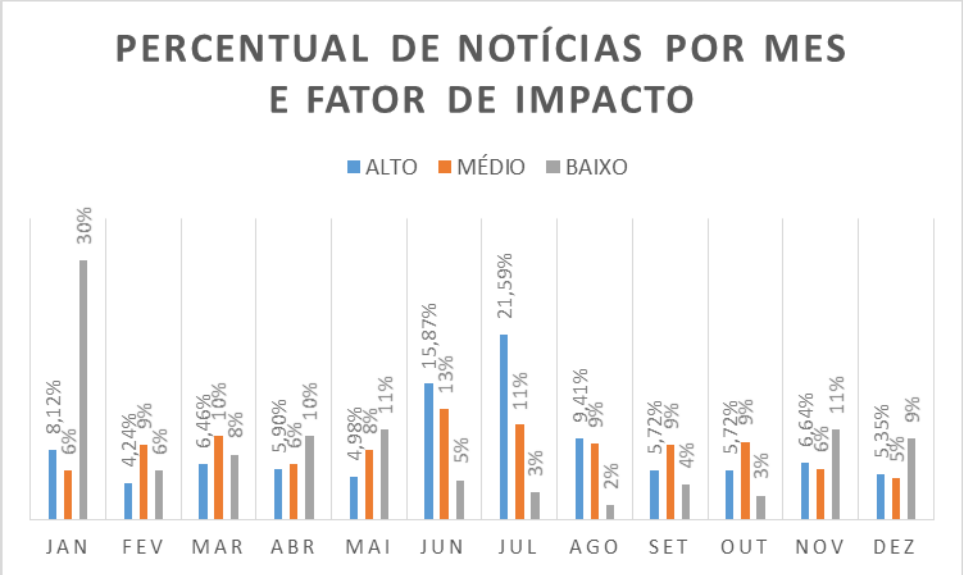


Figura 85 - Percentual de Notícias por mês e Fator de Impacto do Corpus *União Corpora*

Conforme os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por meses presente na união das páginas de fãs, revela o mês de julho com 21,59% de notícias classificadas como fator de impacto alto e em segundo lugar o mês de junho com 15,87%.

É observado também que o maior resultado foi obtido no mês de Janeiro em notícias classificadas como fator de impacto baixo.

5.6.8 Número de mensagens em cada dia da semana por Fator de Impacto do Corpus *União Corpora*

Na figura 86 é apresentada a forma que foi contabilizada a frequência de mensagens por dia da semana presentes nas notícias publicadas de todas páginas de fãs classificada como fator de impacto alto, médio e baixo.

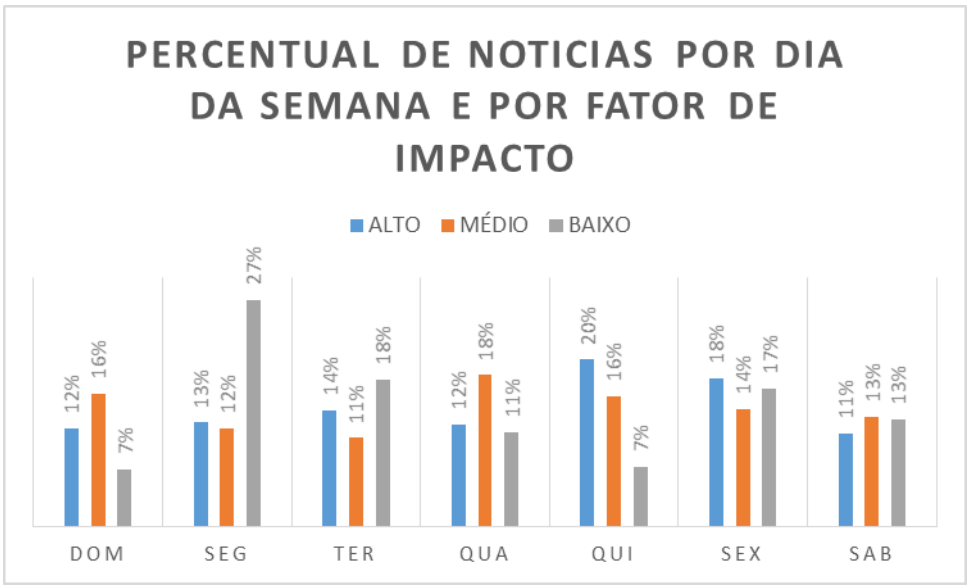


Figura 86 - Percentual de Notícias por dia da semana e Fator de Impacto do Corpus *União Corpora*

Através dos resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por dia da semana presente de todas páginas de fãs, revela que em notícias classificadas como fator de impacto alto o dia da semana vencedor foi quinta-feira.

É observado também que o maior resultado foi obtido segunda-feira em notícias classificadas como fator de impacto baixo.

5.6.9 Número de mensagens em turnos por Fator de Impacto do Corpus *União Corpora*

Na figura 87 é apresentada a forma que foi contabilizada a frequência de mensagens por turno presentes nas notícias publicadas em todas páginas de fãs classificada como fator de impacto alto, médio e baixo.

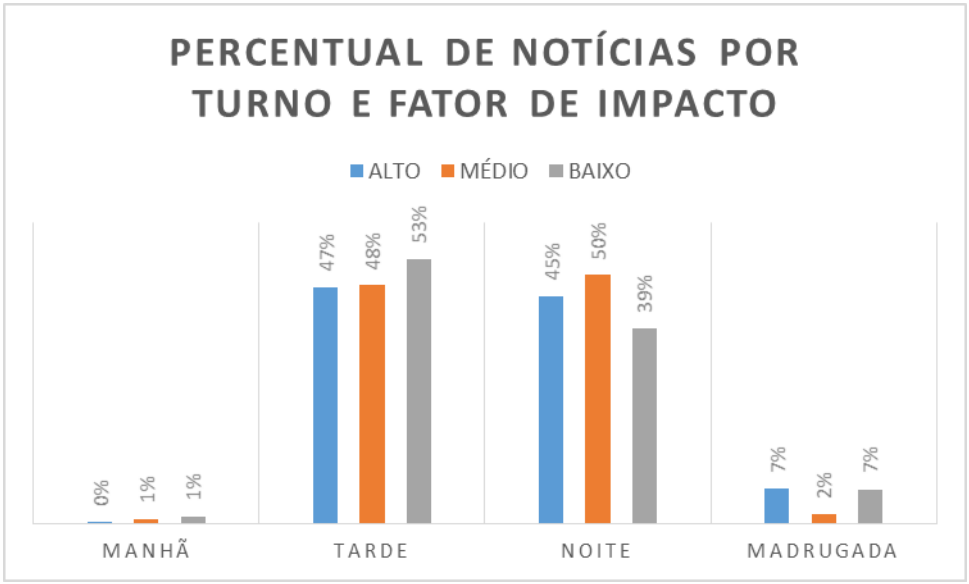


Figura 87 - Percentual do número de mensagens publicadas em turnos por Fator de Impacto do Corpus *União Corpora*

Segundo os resultados apresentados acima pode-se perceber que se tratando de notícias publicadas por turno presente em todas página de fãs, revela que tanto em notícias classificadas como fator de impacto alto, médio e baixo o turno da tarde concentra uma maior quantidade de notícias, e em segundo lugar também do turno da noite.

5.6.10 Número de tipos de mensagens por Fator de Impacto do Corpus *União Corpora*

Na figura 88 é apresentada a forma que foi contabilizada a frequência de tipos de mensagens presentes nas notícias publicadas de todas páginas de fãs classificada como fator de impacto alto, médio e baixo.

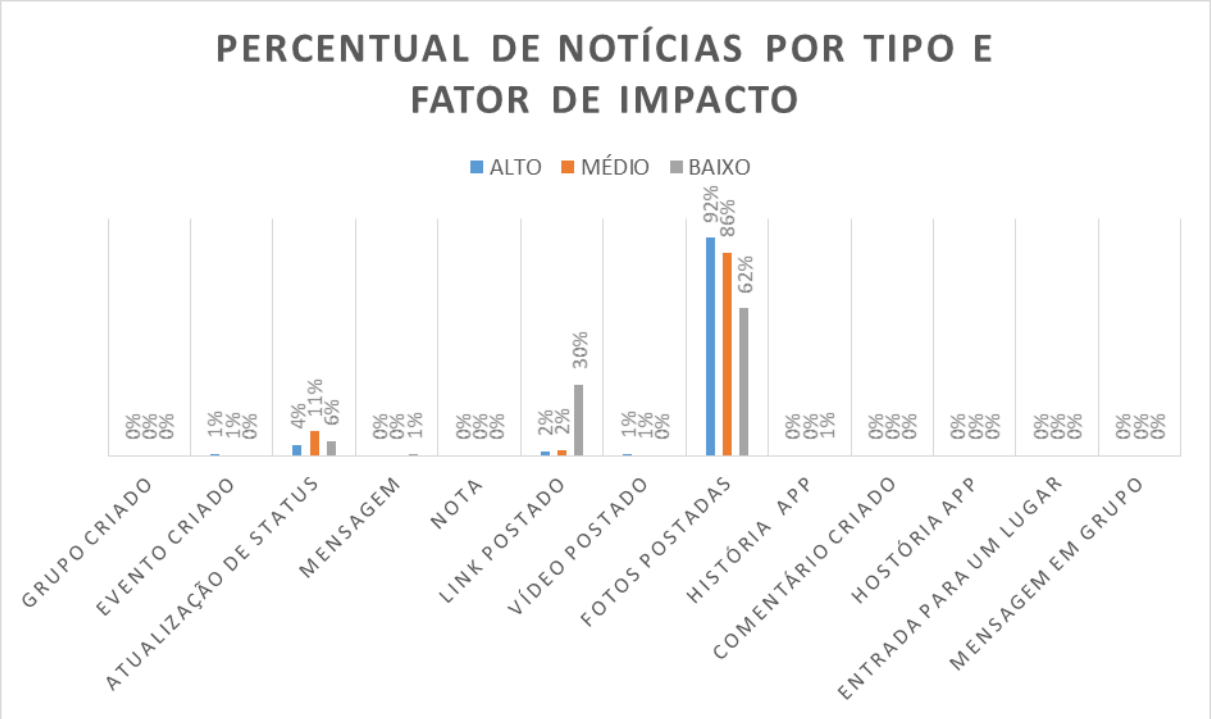


Figura 88 - Percentual de mensagens publicadas em tipos por Fator de Impacto do Corpus *União Corpora*

Consoante os resultados apresentados acima pode-se perceber que se tratando de notícias classificadas como fator de impacto alto, médio e baixo o tipo de notícia que mais ocorre é de Fotos Postadas.

É observado também que o maior resultado foi obtido na classe Fator de Impacto alto com 92% das notícias publicadas pertencem a esse tipo.

5.7 Considerações Finais dos Atributos que exercem influência no Fator de Impacto.

Nesta etapa ao analisarmos estatisticamente a Frequência Relativa, onde FRI é o quociente entre a frequência absoluta do valor do acontecimento e o número total de acontecimentos, conforme representado através da fórmula:

FRI = fi/n

Onde:

n representa o número total de acontecimentos;

fi é o número de vezes que o valor de determinado acontecimento ocorre.

São demonstrados os resultados obtidos através dos dados minerados das cinco páginas de fãs mais acessadas no Brasil, afim de responder as questões desejadas deste trabalho.

5.7.1 Número de palavras por Fator de Impacto

Tabela 94 – Considerações Finais da influência do número de palavras

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Palavras Fa	29,40%	34,77%	30,40%	26,66%	29,35%	34,11%
% Palavras Fm	31,35%	33,98%	38,95%	44,66%	35,43%	54,59%
% Palavras Fb	39,25%	31,25%	30,65%	28,68%	35,22%	11,31%

Em síntese através dos resultados obtidos neste trabalho é identificado que o número de palavras presentes nas mensagens classificadas pela classe Fator de Impacto revela de maneira geral que o número de palavras presentes nas notícias divulgadas na maioria das páginas de fãs utilizadas neste experimento influência no Fator de Impacto.

De tal forma que ao analisarmos estatisticamente a Frequência Relativa, quanto mais palavras utilizadas na composição da notícia divulgada na rede social menor é seu Fator de Impacto.

5.7.2 Número de curtidas, comentários e compartilhamento por Fator de Impacto

Tabela 95 - Considerações Finais da influência do número de curtidas

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Curtidas Fa	76,75%	58,78%	78,47%	74,41%	81,48%	67%
% Curtidas Fm	20,73%	30,80%	21,37%	25,44%	16,70%	31%
% Curtidas Fb	2,52%	10,42%	0,16%	0,15%	1,82%	2%

Tabela 96 - Considerações Finais da influência do número de comentários

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Comentários Fa	75,69%	60,68%	64,53%	69,25%	82,06%	72%
% Comentários Fm	18,66%	29,37%	34,88%	30,54%	13,74%	26%
% Comentários Fb	5,65%	9,95%	0,59%	0,21%	4,20%	2%

Tabela 97 - Considerações Finais da influência do número de compartilhamentos

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Compartilhamentos Fa	91,71%	60,38%	86,51%	66,88%	90,48%	80%
% Compartilhamentos Fm	8,15%	29,89%	13,49%	32,36%	8,80%	20%
% Compartilhamentos Fb	0,14%	9,73%	0,00%	0,76%	0,73%	0%

É identificado também que as interações curtir, comentar e compartilhar estão fortemente relacionadas a classe Fator de Impacto, uma vez que, estas interações compõem a classe Fator de Impacto proposta neste trabalho de tal forma que ao analisarmos estatisticamente a Frequência Relativa é identificado que quanto mais interações, entre as citadas, curtir, comentar e compartilhar maior é seu Fator de Impacto.

5.7.3 Número de Unigram por Fator de Impacto

Tabela 98 - Considerações Finais da influência Unigram1

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Unigram1 Fa	40,93%	39,13%	34,65%	34,94%	37%	43%
% Unigram1 Fm	39,76%	39,13%	34,65%	43,98%	37%	47%
% Unigram1 Fb	19,31%	21,74%	30,70%	21,08%	26%	10%

Tabela 99 - Considerações Finais da influência Unigram2

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Unigram2 Fa	45,09%	42,86%	47,92%	40,80%	27%	34%
% Unigram2 Fm	32,13%	28,57%	45,15%	40,00%	37%	62%
% Unigram2 Fb	22,78%	28,57%	6,93%	19,20%	36%	4%

Tabela 100 - Considerações Finais da influência Unigram3

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Unigram3 Fa	47,73%	30%	40,07%	39,70%	70%	36%
% Unigram3 Fm	22,73%	30%	45,36%	37,48%	20%	61%
% Unigram3 Fb	29,54%	40%	14,57%	20,82%	10%	3%

Tabela 101 - Considerações Finais da influência Unigram4

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Unigram4 Fa	31,42%	20%	48,57%	100%	32%	29%
% Unigram4 Fm	37,94%	40%	50,20%	0%	68%	71%
% Unigram4 Fb	30,64%	40%	1,22%	0%	0%	0%

Tabela 102 - Considerações Finais da influência Unigram5

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Unigram5 Fa	36,86%	33,33%	34,36%	34,78%	0%	33%
% Unigram5 Fm	42,78%	33,33%	45,90%	41,30%	40%	67%
% Unigram5 Fb	20,36%	33,33%	19,74%	23,91%	60%	0%

A técnica Unigram empregada neste experimento, tem a finalidade de utilizar as palavras mais frequentes afim de evidenciar alguma influência de comportamento sobre as classes Fator de Impacto. É apresentado até a quinta palavra mais frequente afim de revelar algum conhecimento.

Observa-se na Tabela 97, que a Unigram1 apresentou percentuais muito equiparados evidenciando que dentro das classes mencionadas não surtiu um efeito que determinasse uma influência da palavra mais frequente para precisar o Fator de Impacto. Na Tabela 98, no que

se refere a Unigram2, observa-se que os resultados ainda que marginais entre as diferentes classes (*Fa e Fm*), que esta técnica atua de forma sutil na determinação do Fa em quatro corpus dos cinco utilizados. Na Unigram3, que corresponde a Tabela 99, identifica-se uma influência de 70% no corpus Lacta para precisar Fa e uma mesma influência porém sutil em outros dois corpus (*Guaraná Antarctica e Garoto*) também na classe Fa. A técnica Unigram4 conforme Tabela 100 observa-se uma influência de 100% no Fa especificamente no corpus de dados minerados da página de fãs da Garoto porém também observa-se que nos demais corpus houve percentuais muito equiparados evidenciando que dentro das classes mencionadas não surtiu um efeito que determinasse uma influência da quarta palavra mais frequente para precisar o Fi. Como também na Tabela 101 , no que se refere a Unigram5, apresentou percentuais equilibrados não decorrendo influência entre as diferentes classes (Fa, Fm e Fb).

5.7.4 Número de Bigram por Fator de Impacto

Tabela 103 - Considerações Finais da influência Bigram1

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Bigram1 Fa	35,60%	33,33%	14,99%	37%	30%	37%
% Bigram1 Fm	34,04%	33,33%	36,90%	43%	54,44%	61%
% Bigram1 Fb	30,36%	33,33%	48,11%	20%	15,56%	3%

Tabela 104 - Considerações Finais da influência Bigram2

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Bigram2 Fa	32,12%	20%	18,41%	38%	28,26%	35%
% Bigram2 Fm	37,48%	40%	69,04%	31%	60,87%	61%
% Bigram2 Fb	30,40%	40%	12,55%	31%	10,87%	3%

Tabela 105 - Considerações Finais da influência Bigram3

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Bigram3 Fa	38,09%	25%	25,79%	52%	31,82%	39%
% Bigram3 Fm	36,32%	50%	51,05%	45%	47,73%	47%
% Bigram3 Fb	25,58%	25%	46,08%	3%	20,45%	14%

Tabela 106 - Considerações Finais da influência Bigram4

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Bigram4 Fa	40,07%	16,67%	35,78%	14%	29,73%	20%
% Bigram4 Fm	40,4%	16,67%	46,08%	64%	51,35%	78%
% Bigram4 Fb	19,53%	66,67%	18,14%	23%	18,92%	2%

Tabela 107 - Considerações Finais da influência Bigram5

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Bigram5 Fa	78,46%	33,33%	26,99%	23%	53,13%	37%
% Bigram5 Fm	21,54%	33,33%	48,88%	55%	28,13%	21%
% Bigram5 Fb	0%	33,33%	21,13%	23%	18,75%	42%

A técnica Bigram empregada aqui neste experimento, tem a finalidade de utilizar pares de palavras mais frequentes afim de evidenciar alguma influência de comportamento sobre as classes Fator de Impacto. Neste experimento é apresentado até o quinto par de palavras mais frequentes com a finalidade de revelar algum conhecimento.

Observa-se na Tabela 102, que a Bigram1 apresentou percentuais muito equiparados evidenciando que dentro das classes mencionadas não surtiu um efeito que determinasse uma influência da palavra mais frequente para precisar o Fator de Impacto. Na Bigram2, que

corresponde a Tabela 103, identifica-se uma influência de 69,04% e de 60,87% nos corpora Hotel Urbano e Lacta para precisar Fm.

Na Tabela 104, no que se refere a Bigram3, observa-se que os resultados ainda que marginais entre as diferentes classes (*Fa e Fm*), que esta técnica atua de forma específica na determinação do Fa em dois corpora dos cinco utilizados (*Guaraná Antarctica e Garoto*). A técnica Bigram4 conforme Tabela 105 observa-se uma influência de 66,67% no Fb especificamente no corpus de dados minerados da página de fãs da Coca Cola porém também observa-se que nos demais corpus houve percentuais muito equiparados evidenciando que dentro das classes mencionadas não surtiu um efeito que determinasse uma influência da quarta palavra mais frequente para precisar o Fi. Na Tabela 106, no que se refere a Bigram5, identifica-se uma influência de 78,46% nos corpus Guaraná Antarctica para precisar Fa, porém dentro das demais classes é identificado percentuais harmônicos não decorrendo influência entre as diferentes classes (*Fa, Fm e Fb*).

5.7.5 Número de Trigram por Fator de Impacto

Tabela 108 - Considerações Finais da influência Trigram1

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Trigram1 Fa	31,80%	29,41%	34,55%	38%	30%	43%
% Trigram1 Fm	38,98%	35,29%	47,56%	41%	54,44%	48%
% Trigram1 Fb	29,22%	35,29%	17,89%	21%	15,56%	9%

Tabela 109 - Considerações Finais da influência Trigram2

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Trigram2 Fa	41,16%	28,57%	24,75%	35%	30%	36%
% Trigram2 Fm	35,20%	28,57%	52,48%	43%	54,44%	61%
% Trigram2 Fb	23,64%	42,86%	22,77%	22%	15,56%	3%

Tabela 110 - Considerações Finais da influência Trigram3

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Trigram3 Fa	36,22%	50,00%	47,26%	65%	31,82%	41%
% Trigram3 Fm	33,19%	33,33%	44,05%	29%	47,73%	37%
% Trigram3 Fb	30,59%	16,67%	8,7%	6%	20,45%	22%

Tabela 111 - Considerações Finais da influência Trigram4

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Trigram4 Fa	30,53%	20%	30,50%	42%	28,26%	34%
% Trigram4 Fm	37,98%	20%	50,82%	33%	60,87%	54%
% Trigram4 Fb	31,49%	60%	13,68%	25%	10,87%	12%

Tabela 112 - Considerações Finais da influência Trigram5

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Trigram5 Fa	36,77%	33,33%	25,72%	13%	28,09%	48%
% Trigram5 Fm	43,39%	33,33%	54,66%	70%	59,55%	52%
% Trigram5 Fb	19,84%	33,33%	19,61%	17%	12,36%	0%

A técnica Trigram empregada aqui neste experimento, tem a finalidade de utilizar o conjunto de três palavras mais frequentes afim de evidenciar alguma influência de comportamento sobre as classes Fator de Impacto. Neste experimento é apresentado até o quinto conjunto de palavras mais frequentes com a finalidade de revelar algum conhecimento.

Observa-se na Tabela 107, no que se refere a técnica Trigram1, identifica-se uma influência de 54,44% especificamente no corpus Lacta para precisar Fm e uma mesma influência porém sutil em outros corpora (*Guaraná Antarctica, Hotel Urbano e Garoto*)

também na classe Fm. Na Tabela 108, apresentou-se uma influência de 54,44%, 52,48% e 43,73% especificamente nos corpora (*Lacta, Hotel Urbano e Garoto*) para precisar Fm e nos demais corpora houve percentuais muito distintos evidenciando que dentro das classes mencionadas não houve influência que determinasse o Fator de Impacto. Na Tabela 109, no que se refere a Técnica Trigram3, observa-se uma influência de 65%, 50%, 47,26 e 36,53% no Fa nos corpora de dados minerados de páginas de fãs (*Garoto, Coca Cola, Hotel Urbano e Guaraná Antarctica*) utilizando o terceiro conjunto de três palavras que mais aparece em todas as notícias. Tabela 110 observa-se uma influência de 60,87% no Fa especificamente no corpus de dados minerados da página de fãs da Lacta e uma influência de 50,82% e 37,98% nos corpora (*Hotel Urbano e Guaraná Antarctica*) na mesma classe. E nos demais corpus percentuais muito equiparados evidenciando que dentro das classes mencionadas não surtiu um efeito que determinasse uma influência da quarta palavra mais frequente para precisar o Fi. Como também na Tabela 111 , no que se refere a Trigram5, apresentou-se uma influência na classe Fm de 70% no corpus Garoto e percentuais equilibrados para a mesma classe nos demais corpora.

5.7.6 Número de Ngram por Fator de Impacto

Tabela 113 - Considerações Finais da influência Ngram5

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Ngram5 Fa	39,11%	34%	36,45%	38,75%	37,38%	40,74%
% Ngram5 Fm	39%	36%	37,94%	40,50%	36,17%	51,66%
% Ngram5 Fb	21,89%	30%	25,61%	20,75%	26,46%	7,60

Tabela 114 - Considerações Finais da influência Ngram10

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Ngram10 Fa	38,05%	30%	21,49%	38,88%	41,04%	41,32%
% Ngram10 Fm	38,87%	34%	24,01%	42,25%	41,43%	49,24%
% Ngram10 Fb	23,08%	36%	50,50%	18,88%	17,53%	9,44%

A técnica Ngram5 empregada aqui neste experimento citada na literatura como sendo o estado da arte de Ngram, tem a finalidade de utilizar o conjunto de cinco palavras mais frequentes afim de evidenciar alguma influência de comportamento sobre as classes Fator de Impacto.

Assim como a Ngram5 o Ngram10 foi utilizado dez conjunto de palavras mais frequentes afim de verificar se há alguma relação de influência na classificação do Fator de Impacto.

Entretanto embora é evidenciado estatisticamente a Frequência Relativa em média de 30% de ocorrência destes conjuntos de cinco e também dez palavras presentes nas notícias, este percentual não é suficiente para identificar a influência no Fator de Impacto entre as classes Fa, Fm e Fb, uma vez que os percentuais se encontra muito próximos de ambas as classes.

5.7.7 Número de mensagens em cada mês por Fator de Impacto

Tabela 115 - Considerações Finais da influência no Mês

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% mês Fa	AGO 22,24%	JAN, JUN, OUT 16%	OUT DEZ 11%	MAR 13%	AGO 28%	JUL 22%
% mês Fm	JAN 11%	FEV 28%	DEZ 23%	JUN AGO 21%	JAN OUT 14%	JUN 13%
% mês Fb	FEV 19%	FEV 28%	JUN 26%	JUL 30%	JAN 26%	JAN 30%

Conforme os resultados apresentados é identificado que não houve uma unanimidade que revele influência no Fator de Impacto de notícias divulgadas em um mês específico.

Entretanto uma relação aqui apresentada neste trabalho como uma proposta que justifique os resultados é que para determinado domínio de notícias o grau de influência pode estar relacionados aos meses que possuem datas comemorativas, como por exemplo o mês de Agosto dia dos pais para as notícias classificadas como Fator de Impacto alto na página de fãs do Guaraná Antarctica, Outubro dia das Crianças para notícias divulgadas nas páginas de fãs da Coca Cola, Hotel Urbano e Lacta, Abril Páscoa mês que antecede o resultado de Março identificado com fator de Impacto da página de fãs da Lacta, Assim como o mês de Janeiro que é confraternização universal identificado também nos resultados da página de fãs da Coca Cola, Guaraná Antarctica, e Chocolate Lacta.

5.7.8 Número de mensagens em cada dia da semana por Fator de Impacto

Tabela 116 - Considerações Finais da influência no Dia da Semana

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Dia da Semana Fa	DOM 16%	SEG 36%	QUI, SEX 16%	QUI 22%	TER 25%	QUI 20%
% Dia da Semana Fm	QUI 17%	QUA 28%	QUA 17%	SEG 17%	SEG 24%	QUA 18%
% Dia da Semana Fb	SEG 20%	QUA 40%	TER 18%	QUA 19%	SEG 24%	SEG 27%

Ao analisarmos a influência do Dia da Semana conforme os resultados apresentados é identificado que assim como nas notícias divulgadas nos meses não houve uma unanimidade que revele influência no Fator de Impacto de notícias divulgadas em um dia da semana específico.

5.7.9 Número de mensagens em cada turno por Fator de Impacto

Tabela 117 - Considerações Finais da influência no Turno

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Turno Fa	NOITE 53%	TARDE 76%	TARDE 47%	NOITE 50%	TARDE 50%	TARDE 47%
% Turno Fm	NOITE 50%	TARDE 76%	NOITE 44%	TARDE 59%	TARDE 49%	NOITE 50%
% Turno Fb	NOITE 50%	TARDE, NOITE 48%	TARDE 67%	TARDE 52%	TARDE, NOITE 41	TARDE 53%

Também ao analisarmos a influência do Turno em que as notícias foram divulgadas é evidenciado que indiferentemente da classe Fator de Impacto a maioria das notícias classificadas como fator de Impacto alto, médio e baixo, todas aparecem como publicadas nos turnos tarde e noite.

Observando mais atentamente dos cinco corpus observados esta característica "Turno" com Fator de Impacto Alto, três corpora identificam o turno da tarde como sendo o mais promissor para alcançar o Fator de Impacto Alto, sendo elas, Coca Cola com 76%, Hotel Urbano com 47% e Lacta com 50%.

5.7.10 Número de tipos de mensagens por Fator de Impacto

Tabela 118 - Considerações Finais da influência por Tipo

	Guaraná Antarctica	*Coca Cola*	*Hotel Urbano*	*Garoto*	*Lacta*	*Base Unida*
% Tipo Fa	Fotos Postadas 62%	Atualização de Status 80%	Fotos Postadas 79%	Fotos Postadas 100%	Fotos Postadas 79%	Fotos Postadas 92%
% Tipo Fm	Fotos Postadas 87%	Atualização de Status 72%	Fotos Postadas 49%	Fotos Postadas 98%	Fotos Postadas 78%	Fotos Postadas 86%
% Tipo Fb	Fotos Postadas 50%	Link Postado 60%	Link Postado 38%	Fotos Postadas 83%	Fotos Postadas 43%	Fotos Postadas 62%

Conforme os tipos de notícias definidos pela Rede Social Facebook, e através dos resultados revelados neste trabalho é identificado que notícias pertencentes a classe Fator de Impacto Alto possuem foto ou imagem presentes em suas divulgações. O que justifique a motivação dos internautas em interagir com as postagens e com essa interação propagar a notícia através das relações pertencentes ao círculo de amigos em comum.

6 PROPOSTA DE MODELO DE PREDIÇÃO

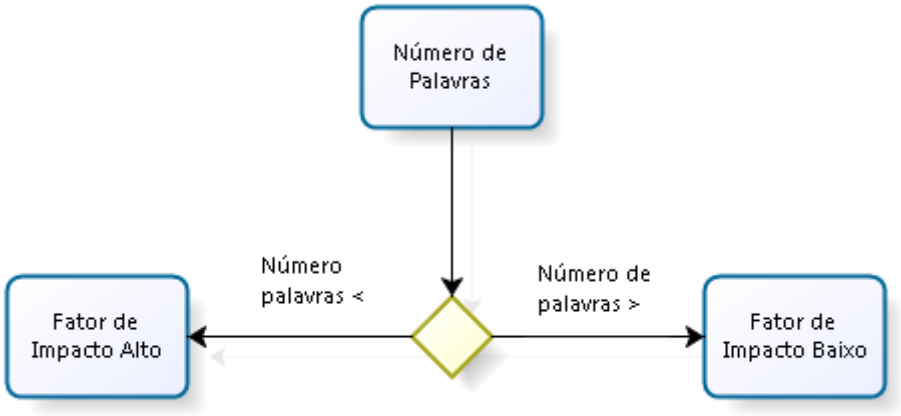
Conforme (CHWIF, 2010) para um sistema não existente é possível construir modelos de simulação de hipotéticos. Um modelo é uma abstração da realidade, que se aproxima do verdadeiro comportamento do sistema, mas sempre mais simples do que o sistema real.

Baseado nas evidências apresentadas no trabalho é proposto um modelo simples que consiste em identificar as características que uma notícia postada numa página de fãs da rede social facebook deve possuir para ser classificada como fator de impacto alto, ou seja alcançar uma quantidade maior de interações.

Embora alguns dos experimentos propostos no trabalho não revelaram influência na classe fator de impacto o modelo proposto identifica as características evidenciadas através da estatística.

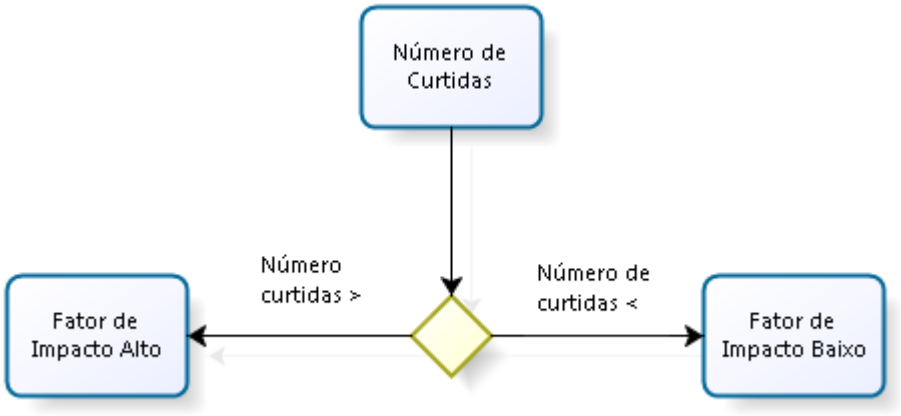
Conforme apresentado na tabela 119, para que uma notícia seja classificada como fator de impacto alto ele deve possuir poucas palavras.

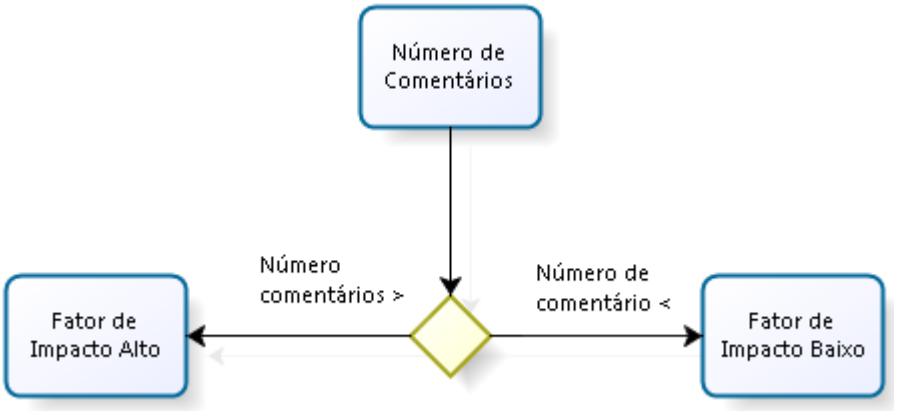
Tabela 119 – Constituição do Modelo de Predição pelo número de palavra

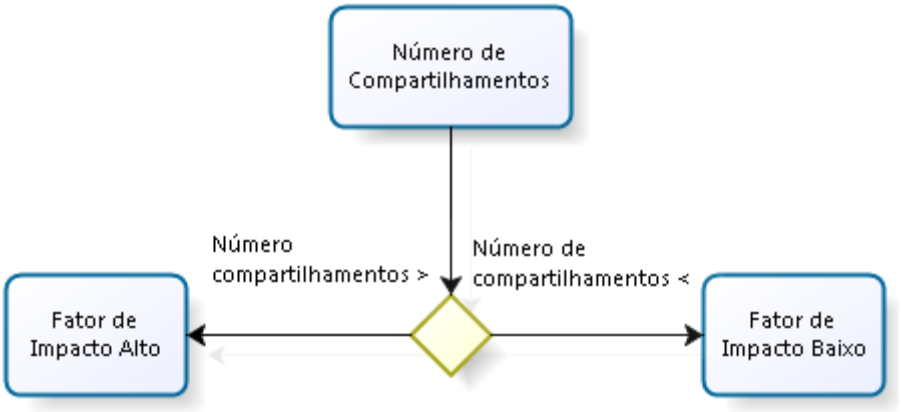
Número de palavras por Fator de Impacto


Conforme apresentado na tabela 120, as interações curtir, comentar e compartilhar estão relacionadas pois foram utilizadas na composição do fator de impacto. Desse modo quanto mais curtidas uma notícia tiver maior é seu fator de impacto, da mesma maneira para outras interações mencionadas, compartilhar e comentar, de modo que quanto mais comentários ou compartilhamentos também maior será o fator de impacto desta notícia.

Tabela 120 - Constituição do Modelo de Predição pelo número de interações

Número de curtidas, comentários e compartilhamento por Fator de Impacto






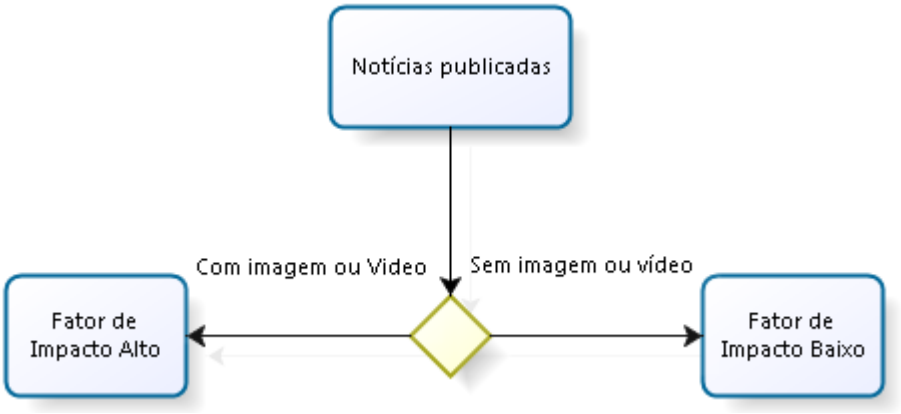
Conforme apresentado na tabela 121, para que uma notícia seja classificada como fator de impacto alto se ela for publicada no turno da tarde maior será a probabilidade de ser classificada como fator de impacto alto.

Tabela 121 - Constituição do Modelo de Predição por turno

Número de mensagens em cada turno por Fator de Impacto

Conforme apresentado na tabela 122, para que uma notícia seja classificada como fator de impacto alto ela deve ser do tipo Foto postada ou Vídeo postado atendendo esta característica maior será a probabilidade de ser classificada como fator de impacto alto.

Tabela 122 - Constituição do Modelo de Predição pelo tipo de notícia

Número de tipos de mensagens por Fator de Impacto.


Desta forma o modelo proposto apresenta características evidenciadas através da estatística encontradas através dos experimentos. Desta forma para que uma notícia atinja um fator de impacto alto, ou seja, um número de interações maior, a notícia deve apresentar no mínimo as seguintes características.

Texto com poucas palavras;

Ser divulgada no turno da tarde;

Apresentar imagem ou vídeo.

O número de interações não foi considerado uma vez que foi utilizado para compor a classe fator de impacto.

7 TRABALHOS FUTUROS

O presente trabalho mostrou a aplicação de algumas técnicas de PLN e também DCBD afim de gerar novos atributos e através deles identificar padrões que influenciam na predição do fator de impacto. Para determinar o fator de impacto foi definido a média das três interações mencionadas, o número de curtidas, o número de comentários e o número de vezes que uma notícia foi compartilhada, com a finalidade de identificar o algoritmo de aprendizagem que melhor prediz a classificação do fator de impacto.

No entanto existem outras alternativas interessantes que também podem ser aplicadas em trabalhos futuros:

- Explorar outras técnicas de PLN como por exemplo a presença de entidades nomeadas ou ainda análise profunda, para verificar se existe a presença de pessoas famosas ou nomes de empresas que influenciam no aumento das interações, ou ainda a ocorrência de classes gramaticais específicas que mais ocorrem.
- Explorar técnicas de análise de sentimento com a finalidade de encontrar indícios de apelos sentimentais influenciando nas notícias com maior interação.
- Explorar o uso de pesos diferenciados para os tipo de interação curtir, comentar e compartilhar, coma a finalidade de, comparar se estes pesos melhoraram ou pioram os resultados da predição deste trabalho.
- Explorar a atribuição de pesos diferenciados às técnicas UNIGRAM, BIGRAM, TRIGRAM e N-Gram, a fim de aumentar o destaque das palavras ou conjunto de palavras que mais ocorrem em uma notícia, uma vez que os resultados aqui apresentados não possibilitaram distinguir evidências que identifique alguma influência de interações através da frequência de palavras ou conjunto de palavras que mais ocorrem em notícias classificadas como fator de impacto alto.
- Por fim, poderia ser usado diferentes algoritmos para a tarefa de classificação, e compara-las com outras abordagens.

CONCLUSÃO

O presente estudo teve como objetivo a aplicação de algoritmos de classificação para análise dos fatores que influenciam na predição do Fator de Impacto em redes sociais. Neste contexto os experimentos aqui realizados, em particular, identificam que dentre as técnicas de classificação com melhor resultado na tarefa de predizer o fator de impacto, o algoritmo J48 apresentou uma taxa de predição de 91,9324% de classificação de instâncias corretas para notícias divulgadas na página de fãs do Guaraná Antarctica e também uma taxa de predição de 86,0963% de classificação de instâncias corretas para notícias divulgadas na página de fãs da Lacta.

Entretanto a mesma Técnica não teve o mesmo sucesso para outros corpora utilizados, o que resultou na identificação de outra técnica de classificação, revelando um desempenho de acertos na utilização do algoritmo DecisionTable para outros três corpus utilizados sendo Coca Cola com 72% de classificação de instâncias correta, Garoto com 85,7658% e também na base de dados Unida com 78,1673% de classificação de instâncias correta conforme (Tabela 91 pág. 75).

Com esses resultados a tarefa de analisar e predizer o impacto de interações em notícias divulgadas em Redes Sociais foi atendida conforme expectativa do trabalho e também que dentre os algoritmos empregados embora o algoritmo J48 apresentou um excelente resultado na predição do Fator de Impacto o algoritmo DecisionTable apresentou um resultado muito bom na maioria dos corpora utilizados.

Outra contribuições deste trabalho foi a definição do Fator de Impacto, que consistem na média das três interações, identificadas como: número de notícias curtidas, comentadas e compartilhadas, e através disso possibilitou a exploração do uso de algoritmos de aprendizado de máquina para predizer o atributo fator de impacto em corpora de páginas de fãs da Rede Social Facebook atendendo as expectativas desta dissertação bem como a utilização de técnicas de Processamento de Linguagem Natural e Descoberta de Conhecimento em base de Dados.

Além disso logo outro resultado esperado é revelar quais são os fatores que exercem influência para que uma notícia divulgada tenha um impacto de interações maior na rede social.

Os resultados indicam que dentre as 5 mensagens do corpus classificadas com maior Fator de Impacto Alto apresentam em média 11 palavras nas mensagens, enquanto, as 5 mensagens classificadas com menor Fator de Impacto Baixo, apresentam em média 33 palavras em cada mensagem, revelando com isso que a quantidade de palavras em uma

notícia postada em uma página de fãs é um padrão que influencia na classificação do fator de impacto. Em outras palavras, o tamanho da mensagem de uma notícia influência uma comunidade de leitores de uma rede social na motivação de interações tanto de forma positiva quanto negativa.

Contudo com emprego de Técnicas de Processamento de Linguagem Natural buscou-se analisar a influência através de junções e combinações de palavras afim de identificar atributos que revelem informações contidas em palavras isoladas ou em conjuntos de palavras, para se determinar o Fa, entretanto os experimentos revelam que ambas as Técnicas (Unigram, Bigram, Trigram e Ngram) apresentam percentuais harmônicos evidenciando que dentro das classes mencionadas não surtiu um efeito que determinasse uma influência dentre as técnicas utilizadas. Embora algumas variações apresentam uma influência em um corpus específico o mesmo não ocorreu em outros corpora utilizados. E de maneira geral estatisticamente os percentuais revelados não apresentam coesão suficiente para identificar a influência no Fator de Impacto entre as classes Fa, Fm e Fb, uma vez que os percentuais se encontra muito próximos de ambas as classes na maioria dos corpora.

É identificado também que o grau de influência do Fator de Impacto Alto em notícias divulgadas por meses do ano, pode estar relacionados aos meses que possuem datas comemorativas para divulgação da notícia proporcionando uma onda de propagação da informação através das interações curtir, comentar e compartilhar.

É revelado também que turno da tarde é o mais promissor para alcançar o Fator de Impacto Alto, de interações.

Outro fator que influencia no Fi é o fato de haver a presença de fotos ou imagens que contribui na motivação dos internautas em interagir com as postagens e com essa interação propagar a notícia através das relações de círculos de amigos.

Os resultados desse trabalho possibilitam também uma orientação para que empresas brasileiras, possam divulgar seus produtos através de notícias mais precisas com oportunidade de alcançar um Impacto Maior para a divulgação de notícias, anúncios e marcas, uma vez que a informação pode ser considerada uma nova classe de ativo econômico, assim como moeda ou ouro.

REFERÊNCIAS

BORGES, E. N., Becker, K., Heuser, C. A., and Galante, R. **A classification-based approach for bibliographic metadata deduplication**. [S.l.: s.n.], IADIS, 2012.

B. WIITHRICH et al.. **Daily prediction of major stock indices from textual www data**. INTERNATIONA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 1998. Proceedings… [S.l.: s.n.], 1998.

CHAVES, Bruno Butilhão. **Estudo de algoritmo AdaBoost de aprendizagem de máquina aplicado a sensores e sistema embarcados**. 2011. Dissertação (Mestrado em Engenharia de Controle e Automação Mecânica) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2011. Disponível em: http:www.teses.usp/disponíveis/3/3152/tde-12062-163740/. Acesso em: 14 abr. 2014.

CHWIF, Leonardo; MEDINA, Afonso C. **Modelagem e simulação de eventos discretos & aplicações**. São Paulo : São Paulo Ed. 2010.

DICIO. **Dicionário online de Português**: definições e significados de mais de 400mil palavras: todas as palavras de A a Z. Disponivel em: http://www.dicio.com.br. Acesso em: 14 dez. 2012.

DIETTERICH, T. Approximate statical test for comparing supervised classification learning algorithms. **Neural Computation**, [S.l.], v.10, n.7, 1895-1924.

FAYYAD, Usama: PIATETSKI-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge form volumes of data. **Communications of the ACM**, [S.l.], v.59, n.11, 27-34, Nov, 1996.

Facebook Query Language. **FQL**. Disponivel em:

https://developers.facebook.com/docs/technical-guides/fql. Acesso em: 01 ago. 2013

GONÇALVES, P. et al. **Comparing and combining sentiment analysis methods**. Proceedings… [S.l.: s.n.], 1-11, 2013.

HAN, J.; KAMBER, M. **Data Mining**: Concepts and Techniques. Nova York: Morgan Kaufmann, 2006.

JURAFSKY, Daniel S. **Speech and Langauge Processing**. [S.l.: s;n.], 1999.

KATTI, FACELi. **Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina.** Rio de Janeiro: LTC, 2011.

KRISTIAN BUZA. **Feedback Prediction for Blogs**. CONFERENCE OF THE GERMAN CLASSIFICATION SOCIETY ON DATA ANALYSIS, MACHINE LEARNING AND KNOWLEDGE DISCOVERY, 36., 2012, Hildesheim, Germany. Proceedings… [S.l.: 1-8.], 2012.

KRISTIAN BUZA. **Feedback Preciction for Blogs**. Department of Computer Science and Information Theory Budapest Univerity of Technology and Economics. Disponível em:<http://www.cs.bme.hu/~buza/pdfs/gfkl_buza_social_media.pdf>.
Acesso em: 20 jan 2013.

LAVRENKO, V. et al. **Language models for _nancial news recommendation**. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 9., 2000. **Proceedings…** [S.l.]: ACM, 2000. p. 389-396

MITTERMAYER, M.A. **Forecasting intraday stock price trends with text mining techniques**. ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 37., 2004. **Proceedings**…Hawaii: IEEE, 2004. p.10.

NAVEGA, Sergio. Princípios Essenciais do Data Mining. In: INFOIMAGE, 2002. [S.l.: s.n.], 2002.

REZENDE, S. O. et al. **Mineração de Dados**, In: REZENDE, S.O. **Sistemas Inteligentes.** [S.l.]: Editora Mnoela Ltda.,2003. p.307-335.

SCHUMAKER, R.P.; CHEN, H. **Textual analysis of stock Market Prediction using breaking news**: The az_n text system. **ACM Transactions on Information Systems,** [S.l.], v. 27, n. 2, 1-29, 2009.

SCHMITT, V. F. **Uma análise comparativa de técnicas de aprendizagem de máquina para prever a popularidade de postagens no facebook**. 2013. 57 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

SCHÜNKE, Marco; BARONE, Dante Augusto. **A Ciência da Antecipação: Um Modelo de Predição de Interações em Redes Sociais**. Conferência IADIS Ibero-Americana WWW/Internet, São Leopoldo, RS, Brasil. **Anais IADIS**, 203-207, 2013.

SCHÜNKE, Marco; DIAS, Letícia. **Analise de Modelos de Predição Baseado em Informações**. Conferência IADIS Ibero-Americana WWW/Internet, São Leopoldo, RS, Brasil. **Anais IADIS**, 165-169, 2013.

SCHOEN Harald et al. **The power of prediction with social media**, **Internet Research**, [S.l], v.23 n.5, 528-543, 2013.

LOH, Stanley. **Lista de stopwords – stoplist (português, inglês, espanhol)**. Disponível em: http://miningtext.blogspot.com.br/2008/11/listas-de-stopwords-stoplist-portugues.html.
Acesso em 11 maio. 2013.

TAN, on.; STEINBACH, M. Kumar. **Introduction to Data Mining.** Pearson Education. [S.l.]: Pearson Education, 2006.

TSYTSARAU, M. PALPANAS, T. Survey on mining subjective data on the web. **Data Mining and Knowledge Discovery**, 24(3):478-514. 2012.