A Sentiment Classification Approach for Books Reviews in Brazilian Portuguese Using Different Feature Extraction Methods

Larissa F. S. Britto[1], Luciano D. S. Pacífico[1]

[1]Departamento de Computação (DC) – Universidade Federal Rural de Pernambuco (UFRPE) – Recife – PE – Brazil

{larissa.feliciana,luciano.pacifico}@ufrpe.br

Abstract. *The huge amount of textual data made available every day on the internet encouraged research in several areas that automatically process and analyze these texts. One of the most popular areas is Sentiment Analysis (SA), but although SA has been a topic widely discussed in recent years, it still faces a shortage of available resources in the Brazilian Portuguese language. This work presents the steps for a complete process of sentiment analysis and classification, from the development of a dataset in Brazilian Portuguese (in the books domain) to the sentiment classification, using some of the main classifiers in the literature and different feature extraction methods.*

Resumo. *A enorme quantidade de dados textuais disponibilizados todos os dias na internet incentivou a pesquisa em diversas áreas que processam e analisam automaticamente esses textos. Uma das áreas mais populares é Análise de Sentimentos, que apesar de ter sido um tópico amplamente discutido nos últimos anos, ainda enfrenta uma escassez de recursos disponíveis para o idioma português brasileiro. Este trabalho apresenta o processo completo de análise e classificação de sentimentos, desde o desenvolvimento de uma base de dados em português (no domínio de livros) até a classificação de sentimentos utilizando alguns dos principais classificadores da literatura e diferentes métodos de extração de características.*

1. Introdução

A popularização da internet e dos smarthphones fez das redes sociais, fóruns e blogs, ricas fontes de dados textuais, onde os usuários costumam expressar todas as suas opiniões, críticas, desejos e emoções. Esses dados são úteis, especialmente, para a Análise de Sentimentos (AS), área responsável pelo estudo computacional dessas opniões. A AS tem se tornado uma das áreas mais populares do Processamento de Linguagem Natural (PLN), devido a suas aplicações de grande impacto, que vão desde análise do desempenho de produtos [Fang and Zhan 2015] à predição de resultados de disputas eleitorais [Jose and Chooralil 2016].

Existem diversos problemas de interesse da AS [Farias and Rosso 2017, Blitzer et al. 2006, Pan et al. 2010], sendo um dos principais a classificação de polaridade [Turney 2002, Zuo 2018]. A classificação de polaridade, é um problema de classificação simples de texto, onde as classes são relativas ao sentimento que está sendo expresso no texto. Entre os classificadores mais utilizados na literatura para essa tarefa, podemos citar o Naive Bayes [Zuo 2018], Máquinas de Vetores de Suporte [Lu and Wu 2019, Guan et al. 2018], Regressão Logística [Al Omari et al. 2019, Ramadhan et al. 2017], Árvores de Decisão e Floresta Aleatória [Rathi et al. 2018, Rane and Kumar 2018, Hegde and Padma 2017]. A classificação de polaridade também é um dos temas mais abordados no português, como em [de Aguiar et al. 2018, Souza and Vieira 2012], onde algoritmos de aprendizagem de máquina foram utilizados para classificar os sentimentos contidos em postagens feitas por usuários de redes sociais, como o Twitter[1]. Em [Oliveira et al. 2019], opinões sobre programas sociais do governo brasileiro, obtidos também em redes sociais, foram analisadas e classificadas com o objetivo de auxiliar a gestão social a nível governamental.

Apesar de toda a popularidade da AS, a língua portuguesa ainda possui um défict de recursos, tais como bases de dados públicas em domínios variados e ferramentas adequadas para o PLN. Esses recursos, que podem ser facilmente encontrados na língua inglesa, são escassos no português brasileiro, o que pode dificultar pesquisas. Para tentar suprir a necessidade por bases de dados, *Web Corpus* (coleção estática de vários documentos baixados da Web) [Schäfer and Bildhauer 2015] têm sido desenvolvidos por pesquisadores. Em [de Souza et al. 2018], um *Web Corpus* composto de comentários sobre hotéis foi desenvolvido. No mesmo trabalho ainda são analisadas ferramentas de pré-processamento de texto existentes quando aplicadas ao idioma português. Um *Web Corpus* contendo comentários sobre a Copa de Mundo de Futebol de 2014 foi desenvolvido em [Moraes et al. 2015], utilizando uma API para o twitter. O Twitter também foi a fonte dos dados em [Brum and das Graças Volpe Nunes 2017], onde o *corpus* TweetSentBR foi proposto e comparados com outros da literatura.

Tendo em vista a escassez de bases de dados textuais em português em alguns domínios, neste trabalho iremos demonstrar cada etapa do desenvolvimento de um *Web Corpus* de comentários sobre livros. Além disso, para atestar a qualidade da base desenvolvida, será feita uma comparação experimental dos classificadores mais comumente utilizados na literatura de AS: Naive Bayes, Árvore de Decisão, Floresta Aleatória, Máquinas de Vetores de Suporte e Regressão Logística. Serão também comparados diferentes métodos de extração de características para texto, como, *Bag-of-Words*, TF-IDF, IDF e extração binária.

As principais contribuições deste trabalho são:

1. Descrição do Desenvolvimento e disponibilização de um *Web Corpus*[2] de comentários de livros em Português Brasileiro;
2. Avaliação do *Web Corpus* desenvolvido, na tarefa de classificação de polaridade;
3. Comparação do desempenho de classificadores e métodos de extração de características aplicados no idioma português.

[1] `www.twitter.com`

[2] `https://github.com/larifeliciana/books-reviews-portuguese`

O trabalho está dividido como segue. Na próxima seção (Seção 2) será apresentado o processo de criação do *corpus* proposto, sendo suas principais características discutidas. Na Seção 3, os classificadores e métodos de extração de características utilizados nesse trabalho serão brevemente descritos. Na Seção 4, os resultados experimentais serão discutidos. Por fim, na Seção 5, as conclusões do trabalho são apresentadas.

2. Base de Dados

Os dados textuais utilizados na AS, são dados que possuem diferentes estruturas e que podem conter diversas irregularidades e ruídos. Por isso algumas etapas são essenciais para a construção de uma base de dados de qualidade. Nesta seção, serão debatidas todas as etapas necessárias para criação de um *Web Corpus* para AS. Essas etapas podem ser vistas na Figura 1.

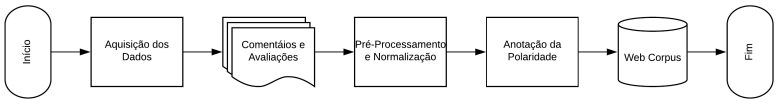

Figura 1. Etapas para o Desenvolvimento do *Web Corpus*.

2.1. Obtenção dos Dados

A obtenção de dados do *website* desejado é o primeiro passo do desenvolvimento do *Web Corpus*. Essa obtenção é geralmente feita de duas formas:

- APIs - Interfaces disponibilizadas pelos próprios *websites* com o intuito de facilitar a interação com desenvolvedores. Nas APIs os dados geralmente estão de forma mais estruturada, o que facilita a sua obtenção;
- *Web Scraping* - Através desta técnica, dados podem ser obtidos dos *websites* automaticamente através de códigos de captura, onde um grande número páginas *web* são simultaneamente examinadas e seus dados extraídos [Vargiu and Urru 2012].

Para este trabalho, foi utilizada a técnica de *Web Scraping*, tendo como alvo páginas de produtos da *Amazon*[3]. Foram extraídos comentários em português sobre livros vendidos nessa plataforma. Além dos comentários, são extraídos também a avaliação do usuário para o produto (classificação feita de 1 a 5). Um total de 2000 comentários foram obtidos.

2.2. Pré-Processamento e Normalização

A maioria dos *websites* como redes sociais, blogs, fóruns e sites de produtos são abertos para o público, isto é, qualquer pessoa pode se cadastrar e inserir seus comentários a respeito dos mais variados temas. Isso pode ser algo positivo tendo em vista a quantidade de documentos textuais disponíveis, porém negativo do ponto de vista da qualidade desses dados, onde os comentários são totalmente desestruturados e podem conter inúmeras irregularidades. Isso faz do pré-processamento uma etapa fundamental do desenvolvimento do *Web Corpus*, onde técnicas de limpeza e normalização de textos são aplicadas com o objetivo de obter um *corpus* padronizado e com menos erros. As seguintes modificações são feitas utilizando a ferramenta NLTK [4] e Expressões Regulares:

[3] `amazon.com.br`

[4] `https://www.nltk.org/book/`

- Conversão de todas as letras para minúsculas;
- Remoção de tags html;
- Remoção de espaços em branco extra;
- Remoção de cacacteres especiais;
- Remoção de *Links* e *Hashtags*.

2.3. Anotação da Polaridade

Na maioria dos *websites* voltados para avaliação de itens, a avaliação é feita em uma escala de 1 a 5, o mesmo ocorre no *Amazon*, onde essa avaliação é simbolicamente representada por estrelas. Através dessas avaliações é feita a anotação da polaridade dos sentimentos contidos nos comentários, isto é, é categorizado se o sentimento expresso naquele comentário é um sentimento positivo ou negativo. Neste trabalho foram consideradas como avaliações positivas, aquelas que receberam 4 ou 5 estrelas e negativas aquelas com 1 ou 2 estrelas. Avaliações neutras foram descartadas. Por fim, o *corpus* é composto por 1000 documentos positivos e 1000 negativos.

2.4. Informações sobre o *Corpus*

Utilizando a biblioteca NLTK, obtemos algumas informações sobre a base de dados proposta neste trabalho. Tais informações podem ser vistas na Tabela 1.

	Positivo	Negativo	Total
Nº Comentários	1000	1000	2000
Setenças / Comentário	4.19	3.47	3.83
Palavras / Comentário	80.35	59.46	69.91
Vocabulário	10186	8381	14567

Tabela 1. Informações sobre o *corpus*

Outra forma de analisar o conteúdo desses comentários é através dos termos mais citados neles. Através desses, podemos observar opniões frequentes dos usuários sobre os itens: pontos que são valorizados pelos mesmos e pontos vistos como negativos. Nas Figuras 2 e 3 podemos ver os unigramas e bigramas que ocorrem com mais frequência no *Web Corpus* proposto, nos comentários negativos e positivos, respectivamente. Para facilitar a visualização de informações importantes, foram removidos termos que continham nomes de autores e de livros.

Numa breve análise nos termos mais frequentes podemos ver alguns pontos que são muito valorizados pelos leitores, tais como, leitura rápida e fácil. Outros pontos importantes a respeito do produto também são considerados pelos compradores, como capa dura e qualidade das folhas. É possível observar também características do serviço de vendas que são importantes para o usuário, como a rápida entrega.

Entre os pontos negativos, podem ser observados incômodos dos usuários em relação a qualidade do material do livro com reclamações em relação a rasgos e amassados. A qualidade da impressão e problemas na entrega do produto também são queixas frequentes.

Figura 2. Unigramas e Bigramas mais frequentes nos comentários positivos.

Figura 3. Unigramas e Bigramas mais frequentes nos comentários negativos.

3. Metodologia

Nesta seção são descritos os métodos de classificação e extração de características utilizados neste trabalho.

3.1. Extração de Características

A extração de características consiste em extrair de textos brutos, as informações que vão alimentar os classificadores. Isso é feito através da transformação do *corpus* obtido, em dados numéricos úteis e suportados pelos algoritmos. Neste trabalho, diferentes métodos para a extração de características foram utilizados:

- *Bag-of-Words* - Um determinado documento de texto é representado como um vetor da quantidade de ocorrência dos seus termos. Nosso *corpus* será então representado por uma matriz semelhante com a da Figura 4.
- *Inverse Document Frequency* - Diferentemente da BoW, o IDF tenta medir o quão importante um termo é. Tentando, assim, diminuir a influência de termos que ocorrem com uma grande frequência, mas que possuem pouca relevância. O escore IDF é dado pela Equação 1.

	bom	livro	legal	amei	eu
Documento 1	1	2	0	1	1
Documento 2	0	1	0	1	1
...	...	...	...	...	...
Documento N	0	1	1	0	0

Figura 4. Exemplo de Bag-of-Words

$$IDF_{t,d} = \log \frac{N}{df_t} \tag{1}$$

Onde t representa o termo e d o documento, N é o número total de documentos, df é o número de documentos em que t ocorre.

- *Term Frequency-Inverse Document Frequency* - Nesse modelo, o escore IDF é combinado com a frequência do termo (Equação 2), que tenta medir o quão relevante um termo é em um determinado documento. O escore TF-IDF é dado pela Equação 3.

$$TF_{t,d} = \frac{f_{t,d}}{\sum_{t_n \in d} f_{t_n,d}} \tag{2}$$

Onde a função f retorna a frequência.

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_{t,d} \tag{3}$$

- Extração Binária - Na extração binária (ou *Bag-of-Words* binária), o vetor númerico contém a representação dos termos que estão ou não presentes no documento [Ziser and Reichart 2016, Britto and Pacífico 2019]. Os valores do vetor de um documento d será dada pela seguinte equação (Equação 4).

$$Bin_{t,d} = \begin{cases} 1 & \text{se } t \in d \\ 0 & \text{senão} \end{cases} \tag{4}$$

3.2. Modelos Selecionados

Esta seção contém uma breve descrição de todos os classificadores utilizados nos experimentos deste trabalho: Naive Bayes, Árvore de Decisão, Floresta Aleatória, Regressão Logística e Máquinas de Vetores de Suporte.

3.3. Árvore de Decisão

A Árvore de Decisão (*Decision Tree* - DT) [Prasad et al. 2015, Bilal et al. 2016, Bayhaqy et al. 2018] é uma estrutura de fluxograma semelhante a uma árvore. A estrutura é útilizada para aplicar um conjunto de regras, onde cada nó interno é responsável por testar um atributo, cada ramificação representa os resultados do teste de determinado nó e cada folha indica o rótulo da classe. Essa árvore é gerada através de processo de divisão, guiado por uma medida de satisfação (a entropia). É utilizada ainda, uma abordagem gulosa para decidir que atributos devem ser levados em consideração para dividir o conjunto de dados em uma determinada iteração do método [Pacifico et al. 2018].

3.4. Floresta Aleatória

O algoritmo de Floresta Aleatória (*Random Forest* - RF) [Pervan and Keles 2017, Parmar et al. 2014] é um método que combina um conjunto de Árvores de Decisão, no intuito de evitar o impacto que ruídos e *outliers* podem ter no resultado de uma única Árvore, o que torna o classificador muito mais robusto. O algoritmo combina diversas Árvores, agregando os votos de diferentes estimadores para decidir a classe final do dado de teste [Criminisi et al. 2011].

3.5. Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (*Support Vector Machine* - SVM) [Ahmad et al. 2017, Fikri and Sarno 2019, Dubey et al. 2019] são algoritmos de aprendizagem supervisionada que se baseiam no princípio de Minimização de Risco Estrutural (*Structural Risk Minimization*) de [Vapnik 1991], com o objetivo de mapear o espaço de características de entrada em um novo espaço onde as classes são linearmente separáveis [Bergsma et al. 2005]. Para isso, o SVM constrói um hiperplano ótimo, que possa separar da melhor forma as instâncias de diferentes classes.

3.6. Naive Bayes

O Naive Bayes (NB) [Yu and Nwet 2020, Abbas et al. 2019] é um modelo probabilístico baseado na aplicação do teorema de Bayes. O NB assume que todas as variáveis são estatisticamente independentes, ou seja, ele desconsidera qualquer correlação e contexto que possa haver entre as palavras de um documento. Para a classificação de uma nova amostra, o classificador calcula a probabilidade condicional dessa nova instância pertencer a cada um das classes, utilizando para isso a probabilidade de cada palavra que compõe o documento.

3.7. Regressão Logística

A Regressão Logística (*Logistic Regression* - LR) [Tyagi and Sharma 2018, Ramadhan et al. 2017] consiste de um modelo discriminativo que é utilizado para a predição da probabilidade das possíveis saídas de uma variável dependente, dado um conjunto de variáveis independentes [Britto and Pacífico 2019]. Para a classificação de sentimentos, é previsto então a probabilidade de uma nova instância conter uma polaridade positiva ou negativa. A Regressão Logística pressupõe que a variável dependente pode ser prevista através da combinação linear das características do problema e dos parâmetros do modelo.

4. Resultados e Discussão

Com o objetivo de avaliar a qualidade da base de dados proposta, alguns classificadores selecionados da literatura de AS foram aplicados, e suas performances comparadas. Além disso, diferentes métodos de extração de características em texto também foram experimentados. Nesta seção, os resultados experimentais para o *Web Corpus* desenvolvido serão apresentados. Foi utilizado um esquema de validação cruzada com 10 *folds*: o *Web Corpus* é dividido aleatoriamente em dez partes (sem sobreposição entre essas partes), e em cada rodada dos experimentos, uma dessas partes é usada como conjunto de teste, enquanto as outras nove partes são utilizadas como conjunto de treinamento dos modelos. Para a avaliação dos modelos, as seguintes métricas foram utilizadas: Acurácia, Revocação (*Recall*), Precisão (*Precision*) e *F-Measure*, as formulas dessas métricas podem ser vistas abaixo (Equações 7 - 8):

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precisão = \frac{TP}{TP + FP} \tag{6}$$

$$Revocação = \frac{TP}{TP + FN} \tag{7}$$

$$F - Measure = \frac{2 \times Precisão \times Revocação}{Precisão + Revocação} \tag{8}$$

onde, TP e TN equivalem a qantidade de documentos positivos e negativos classificados corretamente, e FN e FP a quantidade de documentos positivos e negativos, respectivamente, categorizados de forma errônea. Na Tabela 2 os resultados obtidos podem ser visualizados.

Tabela 2. Resultados Experimentais

		DT	RF	NB	SVM	LR
Acurácia	BoW	0.7185	0.8574	0.876	**0.895**	0.8835
	TFIDF	0.723	0.8685	0.8745	**0.8965**	0.8905
	IDF	0.7185	0.855	0.8765	**0.8915**	0.887
	Binário	0.713	0.8594	0.8745	**0.894**	0.888
Precisão	BoW	0.7271	0.8626	0.8317	**0.891**	0.8819
	TFIDF	0.7263	0.8776	0.8268	**0.8986**	0.8895
	IDF	0.7229	0.8639	0.8303	**0.8915**	0.8854
	Binário	0.7200	0.8671	0.8287	**0.8976**	0.8855
Revocação	BoW	0.704	0.852	**0.944**	0.891	0.886
	TFIDF	0.719	0.858	**0.948**	0.895	0.892
	IDF	0.71	0.843	**0.947**	0.893	0.889
	Binário	0.70	0.85	**0.946**	0.89	0.891
F-Measure	BoW	0.7142	0.8563	0.8840	**0.8941**	0.8834
	TFIDF	0.7219	0.8671	0.8831	**0.8962**	0.8905
	IDF	0.7153	0.8530	0.8846	**0.8916**	0.8870
	Binário	0.7091	0.8579	0.8831	**0.8934**	0.8879

A Máquina de Vetores de Suporte obteve o melhor resultado de acordo com todas as métricas, com exceção da revocação, alcançando quase 90% de acurácia, seguido pela Regressão Logística, que com pouca diferença alcançou até 89.05% de acurácia. Com o pior desempenho se encontra a Árvore de Decisão, alcançando apenas 72.3%. Com

relação aos métodos de extração de características, o TF-IDF obteve o melhor resultado, com uma pequena diferença para os demais.

5. Conclusão

Neste trabalho foi apresentada a descrição detalhada das etapas realizadas para o desenvolvimento do *Web Corpus* no domínio de livros. Devido aos ruídos e erros encontrados em textos produzidos por alguns usuários de *websites*, uma intensa etapa de pré-processamento foi necessária. Através de uma análise no *Web Corpus* desenvolvido, foi possível observar diversos pontos levantados pelos clientes em relação a qualidade dos produtos e do serviço da *Amazon*, alguns desses pontos positivos, como qualidade do material do livro e tempo de entrega, e outros negativos, como problemas de impressão.

Através de uma análise experimental foi comprovada a qualidade da base desenvolvida, tendo os modelos testados obtido resultados condizentes com a literatura da área de AS. Foram comparados o desempenho de alguns dos principais algoritmos utilizados para classificação de sentimentos. Nos testes realizados com a base de dados proposta, Máquina de Vetores de Suporte e a Regressão Logística obtiveram os melhores resultados médios. Entre os métodos de extração de características utilizados, TF-IDF obteve o melhor resultado, porém com uma diferença muito pequena para os outros métodos.

Acreditamos que o *Web Corpus* desenvolvido neste trabalho possa ser utilizado na concepção de outros trabalhos nas mais diversas pesquisas dentro da AS e até mesmo em outros problemas de Processamento de Linguagem Natural. Como trabalhos futuros, pretedemos fazer a aplicação do *corpus* em outras tarefas da análise de sentimentos, como por exemplo o cruzamento e adaptação de domínios, onde *corpora* de diversos domínios são necessários. Esperamos ainda que as etapas para o desenvolvimento deste *Web Corpus* incentive outros pesquisadores a desenvolver e disponibilizar outras bases de dados, contribuindo assim com o avanço das pesquisas relacionadas a dados textuais no português brasileiro.

Agradecimentos

Os autores gostariam de agradecer ao CNPq e à CAPES pelo suporte financeiro.

Referências

Abbas, M., Ali, K., Memon, S., Jamali, A., and Ahmed, A. (2019). Multinomial naive bayes classification model for sentiment analysis.

Ahmad, M., Aftab, S., and Ali, I. (2017). Sentiment analysis of tweets using svm.

Al Omari, M., Al-Hajj, M., Hammami, N., and Sabra, A. (2019). Sentiment classifier: Logistic regression for arabic services' reviews in lebanon. In *2019 International Conference on Computer and Information Sciences (ICCIS)*, pages 1–5.

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., and Kaburuan, E. R. (2018). Sentiment analysis about e-commerce from tweets using decision tree, k-nearest neighbor, and naïve bayes. In *2018 International Conference on Orange Technologies (ICOT)*, pages 1–6.

Bergsma, S., Jung, D., Lau, R., Wang, Y., and Wang, S. (2005). Machine learning approaches to sentiment classification cmput 551 : Course project winter , 2005.

Bilal, M., Israr, H., Shahid, M., and Khan, A. (2016). Sentiment classification of roman-urdu opinions using naïve bayesian, decision tree and knn classification techniques. *Journal of King Saud University - Computer and Information Sciences*, 28(3):330 – 344.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, Stroudsburg, PA, USA. Association for Computational Linguistics.

Britto, L. F. S. and Pacífico, L. D. S. (2019). Análise de sentimentos para revisões de aplicativos mobile em português brasileiro. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 1080–1090, Porto Alegre, RS, Brasil. SBC.

Brum, H. B. and das Graças Volpe Nunes, M. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *CoRR*, abs/1712.08917.

Criminisi, A., Konukoglu, E., and Shotton, J. (2011). Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning.

de Aguiar, E. J., Faiçal, B. S., Ueyama, J., Silva, G. C., and Menolli, A. (2018). Análise de sentimento em redes sociais para a língua portuguesa utilizando algoritmos de classificação. In *Anais do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, Porto Alegre, RS, Brasil. SBC.

de Souza, J. G. R., de Paiva Oliveira, A., and Moreira, A. (2018). Development of a brazilian portuguese hotel's reviews corpus. In *PROPOR*.

Dubey, P., Mishra, A., and Saha, B. K. (2019). Sentiment analysis using svm and deep neural network. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pages 952–957.

Fang, X. and Zhan, J. (2015). Sentiment analysis using product review data. *J Big Data*, 2.

Farias, D. H. and Rosso, P. (2017). Chapter 7 - irony, sarcasm, and sentiment analysis. In Pozzi, F. A., Fersini, E., Messina, E., and Liu, B., editors, *Sentiment Analysis in Social Networks*, pages 113 – 128. Morgan Kaufmann, Boston.

Fikri, M. and Sarno, R. (2019). A comparative study of sentiment analysis using svm and sentiwordnet.

Guan, X., Li, Y., Gong, H., Sun, H., and Zhou, C. (2018). An improved svm for book review sentiment polarity analysis. In *2018 International Conference on Transportation Logistics, Information Communication, Smart City (TLICSC 2018)*. Atlantis Press.

Hegde, Y. and Padma, S. K. (2017). Sentiment analysis using random forest ensemble for mobile product reviews in kannada. In *2017 IEEE 7th International Advance Computing Conference (IACC)*, pages 777–782.

Jose, R. and Chooralil, V. S. (2016). Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble approach. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pages 64–67.

Lu, K. and Wu, J. (2019). Sentiment analysis of film review texts based on sentiment dictionary and svm. In *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, ICIAI 2019, pages 73–77, New York, NY, USA. ACM.

Moraes, S. M. W., Manssour, I. H., and Silveira, M. S. (2015). 7x1-PT: um corpus extraído do twitter para análise de sentimentos em língua portuguesa (7x1-PT: a corpus extracted from twitter for sentiment analysis in Portuguese language). In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, pages 21–25, Natal, Brazil. Sociedade Brasileira de Computação.

Oliveira, D. J. A. S., Bermejo, P. H. d. S., Pereira, J. A. R., and Barbosa, D. A. (2019). A aplicação da técnica de análise de sentimento em mídias sociais como instrumento para as práticas da gestão social em nível governamental. *Revista de Administração Pública*, 53:235 – 251.

Pacifico, L. D. S., Macario, V., and Oliveira, J. F. L. (2018). Plant classification using artificial neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., and Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 751–760, New York, NY, USA. ACM.

Parmar, H., Bhanderi, S., and Shah, G. (2014). Sentiment mining of movie reviews using random forest with tuned hyperparameters.

Pervan, N. and Keles, H. (2017). Sentiment analysis using a random forest classifier on turkish web comments. 59.

Prasad, S. S., Kumar, J., Prabhakar, D. K., and Pal, S. (2015). Sentiment classification: An approach for indian language tweets using decision tree. In Prasath, R., Vuppala, A. K., and Kathirvalavakumar, T., editors, *Mining Intelligence and Knowledge Exploration*, pages 656–663, Cham. Springer International Publishing.

Ramadhan, W. P., Novianty, S. T. M. T. A., and Setianingsih, S. T. M. T. C. (2017). Sentiment analysis using multinomial logistic regression. In *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)*, pages 46–49.

Rane, A. and Kumar, A. (2018). Sentiment classification system of twitter data for us airline service analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 01, pages 769–773.

Rathi, M., Malik, A., Varshney, D., Sharma, R., and Mendiratta, S. (2018). Sentiment analysis of tweets using machine learning approach. In *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–3.

Schäfer, R. and Bildhauer, F. (2015). Web corpus construction roland schäfer and felix bildhauer (freie universität berlin) morgan claypool (synthesis lectures on human language technologies, edited by graeme hirst, volume 22), 2013, 145 pages, paper-bound, isbn 9781608459834, doi:10.2200/s00508ed1v01y201305hlt022. *Computational Linguistics*, 41:161–163.

Souza, M. and Vieira, R. (2012). Sentiment analysis on twitter data for portuguese language. In Caseli, H., Villavicencio, A., Teixeira, A., and Perdigão, F., editors, *Computational Processing of the Portuguese Language*, pages 241–247, Berlin, Heidelberg. Springer Berlin Heidelberg.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tyagi, A. and Sharma, N. (2018). Sentiment analysis using logistic regression and effective word score heuristic. *International Journal of Engineering and Technology(UAE)*, 7:20–23.

Vapnik, V. (1991). Principles of risk minimization for learning theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, pages 831–838, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Vargiu, E. and Urru, M. (2012). Exploiting web scraping in a collaborative filtering- based approach to web advertising. *Artif. Intell. Research*, 2:44–54.

Yu, T. and Nwet, K. T. (2020). Sentiment analysis system for myanmar news using support vector machine and naïve bayes. In Pan, J.-S., Lin, J. C.-W., Liang, Y., and Chu, S.-C., editors, *Genetic and Evolutionary Computing*, pages 551–557, Singapore. Springer Singapore.

Ziser, Y. and Reichart, R. (2016). Neural structural correspondence learning for domain adaptation. *CoRR*, abs/1610.01588.

Zuo, Z. (2018). Sentiment analysis of steam review datasets using naive bayes and decision tree classifier.