An Integrated Approach for Analysing Sentiments on Social Media

Vrinda Tandon, Ritika Mehra*
Department of Computer Science and Engineering, Dev Bhoomi Institute of Technology, India
E-mail: vrinda1804tandon@gmail.com, riti.arora@gmail.com

Keywords: sentimental analysis, natural language processing (NLP), SVC, scikit-learn

Received*:* September 11, 2022

Sentiment analysis is an analytical subfield of Natural Language Processing (NLP) to determine opinion or emotion associated with the body of the text. The requirement for social media sentiment analysis has exceptionally increased with the growing extent of online activities in form of user generated content like posts or comments on social networking platforms. People often share their thoughts, opinions and reviews openly which can further be leveraged to analyze what they feel about a particular topic or their reviews/ feedback about a certain service. This study covers different approaches to conduct social media sentiment analysis on Twitter dataset both balanced and imbalanced obtained from Kaggle. For text analysis, we have implemented various classification techniques such as: Naive Bayes Classification and Support Vector Classification (SVC). It was concluded that SVC on twitter dataset surpassed other classification techniques in terms of performance.

Povzetek: Predstavljena je študija različnih pristopov za analizo sentimenta / razpoloženja na Twitterju, pri čemer se je najbolje izkazala metoda Support Vector Classification.

1 Introduction

Sentiment analysis also referred as emotion AI, helps to determine author's mentality or attitude by classifying their piece of writing as positive, negative or neutral. Sentiment analysis has displayed an intrinsic influence in various domains like product analysis, political campaigns, marketing and competitive research, brand inclination and monitoring. Content posted by the user on social media can be leveraged to make noteworthy interpretation of author's opinion, tone or emotions associated with their posts. The requirement for social media sentiment analysis has exceptionally increased with the growing extent of online activities in form of user generated content like posts or comments on social networking platforms. People often share their thoughts, opinions and reviews openly which can further be leveraged to analyse what they feel about a particular topic or their reviews/ feedback about a certain service.

In social media analysis often known as opinion mining, everything revolves around diving into words to understand the context of the user generated content and the opinions they reveal on such platforms. In this study we have summarized distinct classification techniques on twitter tweets dataset. The objective of this study is to classify sentiments of tweets with and without class weights for both balanced and imbalanced dataset hence, deriving the difference between the results obtained and its impact.

2 Literature review

The data sources on which sentiment analysis can be performed has grown exponentially with time hence; large amount of opinionated data can be fetched from different types of social media and websites. Content posted on different platforms like web-based entertainment platforms for example: movies reviews, social networking platforms or even product reviews can be utilized to make critical translation of author's tone, feelings or feedback related to the text they post. In this section, we have examined the work presented by different researchers on sentiment analysis and briefly discussed their approaches and observations after conducting the study.

Divij et al. in his work covered distinct pre-processing and classification techniques on binary and multi-class movie review dataset. Along with traditional classifiers, modern classifiers like RNN were also implemented. The performance for various approaches was compared where SVM with word embedding was proved to surpass other classification techniques [2]. Naresh et al. in his work determined social media user's opinions by using optimization-based machine learning algorithms. They found that the proposed technique sequential optimization with decision tree provides 89.47% of accuracy compared to other algorithms. Tweets were collected and classified into three categories i.e., positive, negative and neutral. According to the authors, for larger dataset this model will perform faster and will take less time [3]. Mullen et al. conducted a sentiment analysis study on movie review dataset. 1380 movie reviews were collected from a website named epinions.com. The dataset consisted both negative and positive reviews. Support Vector Machine algorithm were applied to train and test the model. For validation purpose three-fold and ten-fold cross validation were used. 84.6% accuracy was obtained using three-fold cross validation whereas 86% accuracy was obtained using tenfold cross validation [1]. The study done to understand different classification techniques on various datasets are tabulated as Table 1.

Table 1: Study to understand different classification techniques on various data sets

Authors	**Paper Title**	**Models /Algorithms**	**Discussion**	**Datasets**	**Year**
Divij Gera, Amita Kapoor [2]	Sentiment Analysis using Scikit Learn: A Review	RNN and BERT models	To perform sentiment analysis on movie reviews	Binary classification dataset from IMDb and multi-class dataset from Rotten Tomatoes)	2022
Naresh, A. and Parimala Venkata Krishna [3]	An efficient approach for sentiment analysis using machine learning algorithm	Sequential minimal optimization with decision tree, Multivariate vehicle regression models	To classify the twitter data.	Airline twitter dataset	2021
Yuxuan Wang, Yutai Hou, Wanxiang Che & Ting Liu [4]	From static to dynamic word representations: a survey	Static and dynamic embedding models	Survey on evaluation metrics and applications of these word embeddings	TOEFL [13], ESL [11], RDWP[14], BM[12], AP and ESSLLI-2008[10]	2020
Kapoor, Amita [5]	Hands-On Artificial Intelligence for IoT: Expert machine learning and deep learning techniques for developing smarter IoT systems.	Machine Learning, Deep Learning and genetics Algorithms	Implement IOT to make their IOT solution Smart	UCI ML (Combined cycle poer plant)	2019
Vanaja, S., & Belwal, M. [6]	Aspect-level sentiment analysis on e-commerce data.	Naïve Bayes algorithm and Support Vector Machine (SVM) algorithm	Aspect-level Sentiment Analysis	Amazon Customer reviews data	2018

Jianqiang, Zhao, and Gui Xiaolin [8]	Comparison research on text pre-processing methods on twitter sentiment analysis	Naive Bayes and Random Forest, Logistic Regression, support vector machine	URLs do not contain useful information for sentiment classification	Stanford Twitter Sentiment Test (STS-Test), SemEval2014, Stanford Twitter Sentiment Gold (STS-Gold), Sentiment Strength Twitter (SS-Twitter) , Sentiment Evaluation (SE-Twitter)	2017
Gamallo, P., & Garcia, M. [9]	A Naive-Bayes Strategy for Sentiment Analysis on English Tweets.	Naive Bayes	For detecting the popularity of English tweets	SemEval2014 organization (tweeti-b.dist.tsv)	2014
Tony Mullen and Nigel Collier [1]	Sentiment analysis using support vector machines with diverse information sources	SVMs based on unigrams and lemmatized versions of the unigram models.	To assign semantic values to phrases and words within a text to be exploited in a more useful way	Epinions.com	2004

2.1 Concern with imbalanced data

One of the major challenges to deal with is imbalanced data. Imbalanced datasets are those datasets in which the observations distribution associated with the target class is not even. In other words, one class label possesses large number of observations as compared to the other class label. The main concern is to accurately and efficiently obtain the likelihood for minority as well as majority class. Imbalanced datasets are prone to give biased results hence to mitigate the issue distinct approaches are utilized.

Model is susceptible to fail when fed poor data, imbalanced data leads to inconsistent results and is considered as one of the major obstacles faced to obtain genuine results. In a study conducted by Alation [16] it was found that more than 80% of the participants were concerned about the quality of the data affecting the progress of their AI executions. https://www.alation.com/blog/alation-sodc-bad-data-spells-trouble-for-ai/

Imbalanced, mislabelled data and data gathered from unknown or non-reliable sources for training and testing tools is the major factor to produce flawed results. Some real-life failure examples induced by flawed data are:

An automated experimental hiring model by Amazon ended up as a failure due to imbalanced training data. The system designed for hiring was found to be biased against women candidates and trained itself by inferring male candidates better.

A racial inclination was found in health prediction algorithms used by US hospital and insurance organisations. The study published in science unveiled that the algorithm was found to recommend white patients over Black patients.

A predictive tool to identify covid-19 and diagnose patients was found to be not fit for clinical use by its own researchers. Derek Driggs' group observed that the trained dataset consisted scans of patients in lying and standing positions which inferred patients in lying position as seriously ill. The algorithm to identify covid-19 risk was inefficient as it was solely giving results based on the position of patients scanned.

In case of imbalanced dataset, the chances of algorithm being biased to the majority class are quite high and the main objective becomes to mitigate misclassification by minority class by setting a higher-class weight to minority class and simultaneously lowering the class weight to majority class. In this study, different weights were assigned to classes to improvise the performance for both binary and multiclass imbalanced data.

3 Working flowchart of proposed work

We implemented various classification techniques for text classification like: Multinomial Naïve Bayes, Bernoulli NB and SVC on the tweets posted by the user. The algorithm and working flowchart (Figure 1.) for the proposed work is as follows:

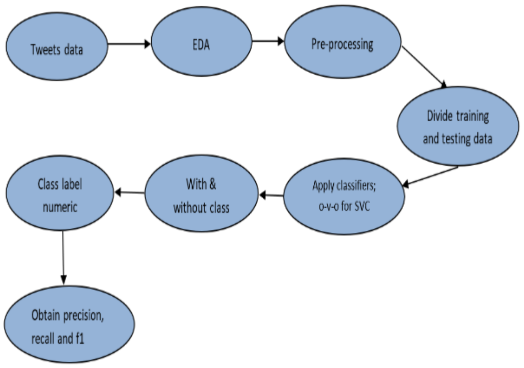

Tweets data
EDA
Pre-processing
Divide training and testing data
Apply classifiers; o-v-o for SVC
With & without class
Class label numeric
Obtain precision, recall and f1

Figure 1: Proposed working flow chart.

3.1 Algorithm for proposed work

Step 1. Divide data into training and testing in 80 and 20 proportions.
Step 2. Determine nature of the dataset i.e., balanced or imbalanced using seaborn. countplot () for sentiment distribution.
Step 3. Apply Naive Bayes Multinomial, Bernoulli classifier and SVC classifier using o-v-o approach as per the nature of the dataset.
Step 4. Predict the test dataset using predict ().
Step 5. Classify tweets with numerical labels.
Step 6. Determine f1-score, precision and recall for each classifier.

4 Dataset utilised

Determining sentiment score is one of the prominent approaches to access emotion or tone of the text. This scaling system assigns scores corresponding to the tone of the text i.e., positive, negative or neutral making it easier to understand. We have utilized twitter Twitter tweets Sentiment Dataset obtained from kaggle to classify tweets sentiment. This multiclass dataset consisted of four columns i.e., textID, text, selected_text and sentiments associated with the tweets. The sentiments were labelled as positive, neutral and negative. In total there were 27.5k tweets present in the dataset.

4.1 Data pre-processing

The sentiments column was originally labelled as neutral, positive and negative so, in order to get sentiment distribution of the dataset we replaced neutral, positive and negative labels with numerical values 0, 1 and -1 respectively. The data containing 27.5k tweets was split into three numerical categories -1 to 1 from negative to positive sentiments associated with the tweets. And the sentiment distribution (Figure 2) of the dataset was represented using seaborn. countplot to draw the ordinal positions on the axis. This was done after importing the seaborn module in collab.

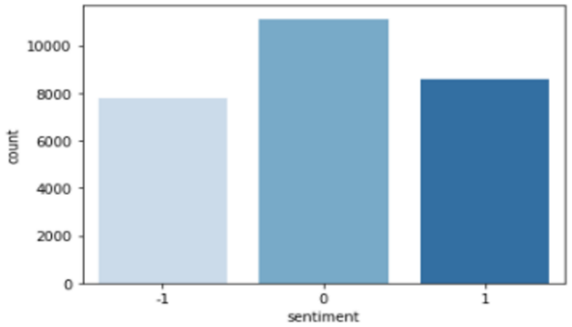
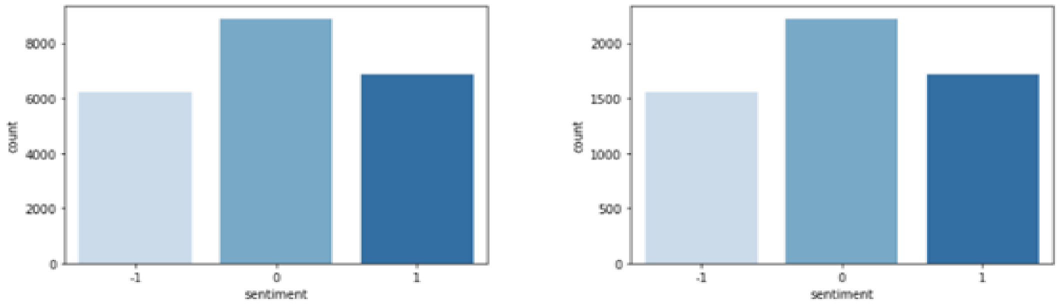


Figure 2: Sentiment distribution of the twitter data.

na values were present in the dataset were removed using fillna (). All the numerical values were also removed from the textual dataset followed by the punctuation removal for pre-processing the data. Conversion to lowercase and expanding contractions were also performed. The dataset was then divided into training and testing data (Figure 3) where plotting was done in order to ensure stratified data split.



Training dataset

Testing dataset

Figure 3: Training and testing dataset.

4.2 Imbalanced multiclass dataset

Imbalanced datasets are those datasets in which the observations distribution associated with the target class is not even. In other words, one class label possesses large number of observations as compared to the other class label. The main concern is to accurately and efficiently obtain the likelihood for minority as well as majority class. Imbalanced datasets are prone to give biased results hence to mitigate the issue distinct approaches are utilized. After obtaining the summary of the dataset using info() method similar operations were performed. The columns were labeled as -1, 1 and 0 for negative, neutral and positive labels after which seaborn_countplot was utilized to represent sentiment distribution as represented in the Figure 4. Numerical values as well as punctuations were removed; lowercase conversion was also done for pre-processing the data.

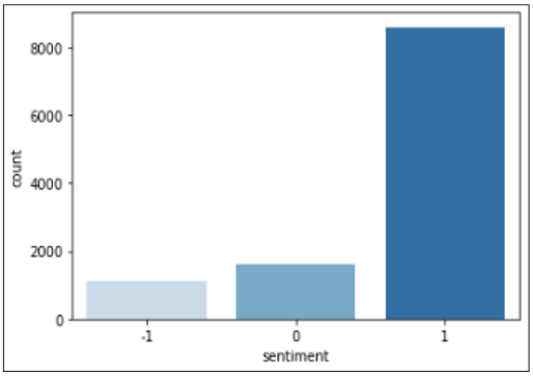
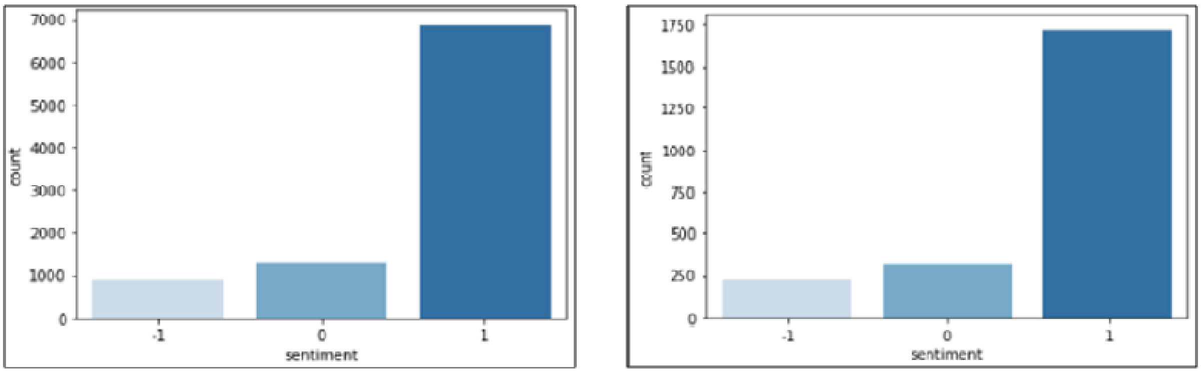


Figure 4: Sentiment distribution of the imbalanced twitter data.

This imbalanced dataset was divided into training and testing data (Figure 5) where plotting was done in order to ensure stratified data split.



Training dataset

Testing dataset

Figure 5: Training and testing dataset

5 Classification techniques implemented

5.1 Naive-Bayes classifier

Naive Bayes is a supervised learning algorithm based on Bayes Theorem. This probabilistic machine learning algorithm predicts on the basis of the probability of an object and is used to solve classification problems. Bayes Theorem also known as Bayes law is a mathematical formula for determining conditional probability as:

$$P(A/B) = \frac{P(B/A)\ P(A)}{P(B)}$$

The fundamental Naive Bayes assumption is that each feature each holds independent and equal contribution to the outcome. The three types of naive Bayes models are:

Gaussian Naive-Bayes classifier**:** This model assumes a normal distribution of features when working with continuous data and likelihood of the features is given as:

$$p(x=v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

Multinomial Naive-Bayes classifier: As the name suggests, this classifier is utilized when we have multinomial distributed data. This is specifically used for document classification and conditional probability formula is given as:

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} {p_{ki}}^{x_i}$$

Bernoulli Naive-Bayes classifier: It is a multivariate event model also popular for document classification where binary term occurrence features are used instead of term frequencies. Here features are independent booleans describing inputs. The likelihood of features is given as:

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1 - p_{ki})^{(1-x_i)}$$

5.2 Naive-Bayes classifier on the data

Data cleaning was done followed by stop-words removal. In order to display frequently occurring features in the corpus a WordCloud (Figure 6) was also produced. The dataset was then vectorized using TF-IDF vectorizer.

The Multinomial and Bernoulli Classifiers provided by Scikit-Learn library were trained and validated using the obtained vectors.

get
good
love
lol
day
like
im
go
got
work
dont
today
going
cant
happy
think
know
time
one
back
home
really
night
want
mothers
morning
well
miss
thanks
tomorrow

Figure 6: WordCloud for twitter dataset.

In case of imbalanced dataset, the chances of algorithm being biased to the majority class are quite high and the main objective becomes to mitigate misclassification by minority class by setting a higher-class weight to minority class and simultaneously lowering the class weight to majority class. Therefore, different weights were assigned to classes to improvise the performance.

5.2 Splitting multiclass classification into binary classification

The algorithms designed for binary classification cannot be leveraged in multi-class classification problems so, to mitigate this issue we use heuristic methods like one-vs-rest and one-vs-one methods to make binary classifiers work as multiclass classifiers.

One-Vs-Rest Classification Model: also known as one-vs-all is a heuristic method to enable binary classification algorithms work as multi-class classification algorithms. In this technique the multi-class data is split as binary classification data in order to apply binary classification algorithms so that it can be ultimately converted into binary classification data [15].

One-vs-One Classification model: this approach is similar to o-v-r as it also functions by splitting the data i.e. by splitting multi-class dataset into binary classification problem. The primary difference between o-v-r and o-v-o is that this classification model groups dataset into one single data file as an opponent to every other class [15].

In our dataset we used one-vs-one (ovo) classification strategy in order to split multi-class classification into binary classification problem per each sets of classes. We trained the underlying classifier as it is after which class weights were presented. Adjusted class weights were utilized to train the data.

6 Results

After completing Sentiment Analysis on twitter's dataset, testing and validation on various classifiers used in the study was done by determining the F1 score, precision and recall. It was evident in this study that the versions of Naive Bayes Classifier gave inaccurate results for imbalanced dataset which were then removed by introducing weights to the class. The results after conducting the study for balanced and imbalanced multiclass dataset are displayed in Table 2. Results for each algorithm for balanced multiclass dataset with and without weights were presented in which no major variation after introducing class weights was observed due to balanced nature of the dataset. SVC displayed better results.

Table 2: Observation Table F1-score for balanced dataset

S No.	Classifier	F1-score
1.	Bernoulli Naive-Bayes (without classweights)	[0.5473 0.6598 0.6851]
2.	Multinomial Naive-Bayes (without class weights)	[0.4914 0.6555 0.6280]
3.	Bernoulli Naive-Bayes (With class weights)	[0.4955 0.6371 0.5163]
4.	MultinomialNaive-Bayes (with class weights)	[0.4292 0.6356

		0.3952]
5.	Support Vector Classifier (without class weights)	[0.6683 0.6944 0.7489]
6.	Support Vector Classifier (with class weights)	[0.6803 0.6738 0.7547]

Inaccurate results for each algorithm were found in case of highly imbalanced dataset so, to mitigate the issue class weights were presented after which SVC and Bernoulli displayed improved results (Table 3).

Table 3: Observation Table F1-score for imbalanced dataset

S.No.	**Classifier**	**F1-score**
1	Bernoulli Naive-Bayes (without classweights	0.6608
2	Multinomial Naive-Bayes (without class weights)	0.6560
3	Bernoulli Naive-Bayes (with classweights)	0.6880
4	Multinomial Naive-Bayes (with class weights)	0.6714
5	Support Vector Classifier (without class weights)	0.7024
6	Support Vector Classifier (with class weights)	0.7446

7 Conclusion and future scope

In this study we covered different approaches to conduct social media sentiment analysis on Twitter tweets dataset fetched from Kaggle. The twitter dataset was multiclass data which required more data pre-processing and hence was closer to the dataset in real life situation where sentiment analysis is conducted. Binary classification algorithms were leveraged in multiclass dataset with the help of o-v-o heuristic technique. These tweets were classified in positive, negative and neutral categories using distinct classification approaches. We implemented various classification techniques using Scikit-learn library for comparative analysis like Bernoulli Naive-Bayes, Multinomial Naive- Bayes and SVC using TF-IDF vectorizer.

A model is susceptible to fail and generate inaccurate results when fed poor i.e. imbalanced data. Highly imbalanced data can create a huge impact on the model's performance and in real life situation it is not surprising to encounter unbalanced datasets. Therefore, it is very important to select the right evaluation matrix in such scenarios. In our study we have utilised F1 score as our evaluation matrix and class weights were also introduced to obtain improved results and enabled us to study Multinomial, Bernoulli and Support Vector classifier with both class weights and without class weights for balanced and imbalanced dataset. For balanced dataset, no major variation was observed after introducing class weights while for imbalanced dataset, the results improved significantly where the Support Vector Classifiers (SVC) ended up being the best performing classifier with class weights. Further, we are focused to test GloVe word embeddings along with different approaches to handle imbalanced data for sentiment analysis.

References

[1] Mullen, Tony, and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. International conference on empirical methods in natural language processing, 412-418. 2004. https://doi.org/10.3115/1219044.1219069

[2] Divij Gera, D., & Kapoor A. Sentiment Analysis using Scikit Learn: A Review, 2022. https://doi.org/10.13140/RG.2.2.26189.10720

[3] Naresh, A. and Parimala Venkata Krishna. An efficient approach for sentiment analysis using machine learning algorithm. Evolutionary Intelligence. 14, 725- 731, 2021. https://doi.org/10.1007/s12065-020-00429-1

[4] Wang, Y., Hou, Y., Che, W., & Liu, T. From static to dynamic word representations: a survey. International Journal of Machine Learning and Cybernetics, 11(7),1611-1630, 2020. https://doi.org/10.1007/s13042-020-01069-8

[5] Kapoor Amita, Hands-On Artificial Intelligence for IoT: Expert machine learning and deep learning techniques for developing smarter IoT systems. ISBN-978-1-78883-606-5, 2019, Packt Publishing Limited, Birmingham, UK.

[6] Vanaja, S., & Belwal, M. Aspect-level sentiment analysis on E-commerce IEEE International Conference on Inventive Research in Computing Applications (ICIRCA), 1275-1279. 2018. https://doi.org/10.1109/icirca.2018.8597286

[7] Kumar, Y., Sharma, H., & Pal, R. Popularity Measuring and Prediction Mining of IPL Team Using Machine Learning. IEEE 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) ICRITO, 1-5, 2021. https://doi.org/10.1109/icrito51393.2021.9596405

[8] Jianqiang, Zhao, and Gui Xiaolin. Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access 5: 2870- 2879, 2017.
https://doi.org/10.1109/access.2017.2672677

[9] Gamallo, P., & Garcia, M. Citius: A Naive-Bayes Strategy forSentiment Analysis on English Tweets. Semeval@ coling, *pp.* 171-175, 2014.
https://doi.org/10.3115/v1/s14-2026

[10] Baroni M, Evert S, Lenci A. Bridging the gap between semantic theory and computational simulations. In: Proc. of the ESSLLI workshop on distributional lexical semantic, 2008.
https://archive.illc.uva.nl/ESSLLI2008/Materials/BaroniEvertLenci/BaroniEvertLenci.pdf

[11] Baroni M, Murphy B, Barbu E, Poesio M. Strudel: a corpus-based semantic model based on properties and types. Cogn Sci 34:222–254, 2010.
https://doi.org/10.1111/j.1551-6709.2009.01068.x

[12] Landauer TK, Dutnais ST. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol Rev 211–240, 1997.*
https://doi.org/10.1037/0033-295x.104.2.211

[13] Turney PD. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: De Raedt L, Flach P (eds) Machine learning: ECML 2001., Springer, Berlin Heidelberg, 491–502, 2001.
https://doi.org/10.1007/3-540-44795-4_42

[14] Jarmasz M, Szpakowicz S. Roget’s Thesaurus and semantic similarity. In: Proc. of RANLP, pp 21, 2003.
https://doi.org/10.1075/cilt.260.12jar

[15] Multiclass classification:
https://www.analyticsvidhya.com/blog/2021/05/multiclass-classification- using-svm/

[16] Alation study:
https://www.alation.com/blog/alation-sodc-bad-data-spells-trouble-for-ai/