Sentence Classification using Machine Learning with Term Frequency–Inverse Document Frequency with N-Gram

Nagendra Nagaraj, Chandra J

Department of Computer Science, Christ University, Bangalore, India

Corresponding author: Nagendra Nagaraj, Email: nagendra.n@res.christuniversity.in

Automatic text classification has proven to be a vital method for managing and processing a very large text area—the volume of digital materials that is spreading and growing on a daily basis. In general, text plays an important role in classifying, extracting, and summarizing information, searching for text, and answering questions. This paper demonstrates machine learning techniques are used for the text classification process..And also, with the vast rapid growth of text analysis in all areas, the demand for automatic text classification has widely improved by day by day. The pattern of text classification has been the subject of a lot of research and development works in recent times of natural language processing is a field that entails a lot of work. This paper represents a text classification technique using the term frequency-inverse document frequency and N-Gram. Also compared the performances of a different model. The recommended model is adopted with four different algorithms and compared with generated results from the algorithms. The linear support vector machine is most relevant to this work with our proposed model. The final result shows a significant accuracy compared with earlier methods.

Keywords: Natural Language Processing, TF-IDF, Data Mining, Classification, N-Gram, Machine Learning.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 337–346. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-35

1 Introduction

Everyone's life is now tightly entwined with technology. Technology has drastically increased almost every aspect of modern life, from phone conversations to satellite launches into space [1-4]. The ability to create and handle information, in general, has influenced technological advancement. Every day, an average of 1,829 petabytes are processed over the internet, according to the US national security agency [5-6]. Regulating and facilitating the flow of data and information communicated over the internet has become required due to the rapid growth of data and information conveyed over the internet [7-8]. Several commercial and social applications have been introduced to this end. Security, research, and sentiment analysis are all components of data and information that can significantly benefit enterprises, governments, and the general public [9-10]. Several customized strategies aid us in a variety of activities, including classification, summarising, and swiftly obtaining and managing data [11-12]. Machine learning (ML) and deep learning (DL) algorithms are just a few examples of algorithms that can be used to handle data [13-14]. There is, nevertheless, a significant amount of material available. Computer techniques can aid in the processing of information from top to bottom, as well as the analysis of complete manuscripts as well as individual words [15-16].

Human-generated 'natural' data in the form of texts, audio, video, and other formats is quickly rising in the actual world. Text analysis has sparked interest in approaches and technologies that can automatically extract useful information from massive amounts of unstructured data [17]. Text mining, a hybrid of techniques including data mining, machine learning, and computational linguistics, is one of the most important methods. Text-mining [18-19] is the process of extracting information and patterns from textual data. Manual text mining is a modest technique to text mining in which a human scans the text and looks for helpful information. Automatic text mining is a more analytical technique in terms of speed and cost [20-22].

Unstructured textual data have expanded rapidly in the finance industry [23]. Where the text mining has a lot of potentials. [24] analyze various operations in the financial domain in which text mining could play an important role. It has numerous applications in this industry, such as several predictions, customer relationship management (CRM), and cybersecurity issues. In recent years, many novel methods have been proposed for analyzing financial results, and artificial intelligence has made it possible to analyze and even predict economic outcomes based on historical data.

The processing of structured or semi-structured data becomes complicated in all fields due to excessive data [25-26] from different domains. Many techniques and algorithms can be used to understand the data. Still, this study will focus on one of them to compare with various models, providing better result or accuracy with the text data set. And using term frequency-inverse document frequency with n-gram. Term frequency-inverse document frequency is a numerical statistic representing the importance of keywords for specific sentences or documents. It can be used to provide keywords that can be delivered to describe or categorize certain records. For example, an article writer runs a blog with thousands of contributors and has employed an intern whose main job is to add new blog posts every day and every week. And it has been observed that detainees often do not pay attention to the tags, so many articles' posts are not categorized in the articles he posted. This is one of the ideal conditions for using the term frequency-inverse document frequency algorithm to identify the category for the article using machine learning, which can automatically identify articles. And saves bloggers and interns a lot of time as they don't have to worry about tags or category [27].

There are many applications or operations of text classification in the real world. Topics generally categorize news stories, categories often category content or products. Users can be classified into the group by discussing a product or brand online. Text mining encompasses data mining, information retrieval from multiple sources, using natural language processing techniques. Thus, the need to

effectively groping, clean, and classify the data is growing in an exponential. One such methodology to identify the text data to understand from different sources is the term frequency-inverse document frequency model to solve the text data issue.

The combined weight of two statistics, term frequency and inverse document frequency, is called term frequency-inverse document frequency. The frequency of words in a document or corpus is measured by term frequency. Each time is assigned a weight based on the number of occurrences in the document 'd' denoted by tf, t, d, and supplied by the following expression (1).

$$tf(t,d) = 0.5 + \{0.5\, xf(t,d)\} / \max\{f(t',d): t' \in d\} \quad (1)$$

The number of text keywords containing the term 't' is reproduced by document frequency. df(t) = occurrence of 't' in documents using inverse document frequency, which is the inverse of document frequency. idf(t) = (N/df) idf(t) = (N/df) idf(t) = (N/df) idf(t) = (N/df) idf(Where 'N' is the number of document's in the collection.

Most text classification documents, articles, and online content courses, on the other hand, use binary (one or zero) text classification techniques like email spam filtering (spam vs. ham) and sentiment text analysis (positive vs. negative vs. neutral) in the provided text from the sources. The real world, in many circumstances, is far more difficult than that. As a result, the wording of today's phrase will categorise consumer financial concerns into pre-defined classes for each class.

2 Related Work

One area of exploration research around the archives as indicated by their source or source style, with measurably recorded expressive varieties [28] filling in as a significant aide. Models are creator, distributer (e.g., The New York Times versus The Daily News), local speaker foundation, and "forehead" (e.g., "famous" versus crude) [29-30].

Determining text type is another related field of research; subjective genres such as "editorial" are frequently used categories [31]. Other studies [32] look for characteristics that clearly demonstrate the use of subjective language. While categorization and subjectivity recognition algorithms aid in the identification of papers that communicate an opinion, they fall short of the goal of identifying that opinion.

The majority of earlier sentiment-based text classification research was at least partially knowledge-based. Some of the research focuses on text classification of individual keywords or phrases' semantic coordination utilising linguistic or a pre-selected collection of seed keywords [33]. The manual or semi-manual building of discriminant-word lexicons, as well as models influenced by cognitive linguistics from natural language processing, have all been used in the work on text sentiment-based classification of complete document sentences [34-35]. Humans may not always have the best ability to categorise terms for discrimination. [36] work on categorising reviews based on their text. Based on the information between document sentence phrases and the terms "great" and "bad," the author created a special unsupervised learning technique. The information is derived from statistics obtained from a search engine. To identify the task's features of complexity, we use multiple entirely prior-knowledge supervised machine learning algorithms.

The author first suggests emotional text classification using the machine learning technique. Analyzed naive bayes, max entropy, and support vector machine technique models to analyze unigram and bigram data [37]. A support vector machine paired with unigrams delivered the best result in their experiments.

The author performed a sentiment text classification using a support vector machine and collected data from multiple sources [38]. The work showed that using a hybrid support vector machine with extracted features based on the theory of Osgood's gives the best results for the classification. This technique worked well but considered contextual text classifications, and the overall impact was significantly reduced due to field variability. The accuracy achieved with the proposed method was 86.6%.

The author created a combinational model technique to examine the linguistic characteristics of the review articles from online sources [39] to assess their usefulness. The support vector machine (SVM) algorithm is used for text classification. The quality of the review is good if it accommodates both subjective and objective information. However, the analysis efficiency was only 71% due to a fuzzy search technique for opinion mining on text, which created a significant problem when a misspelled vital word was found. The sentiment analysis was lugged out on twitter messages using various text classification functions using the n-gram feature and the lexicon feature, part of the speech feature. Work was mainly subject-specific and accomplished an accuracy of around 82%, and they concluded that the part of speech feature reduced the accuracy level [40]. The author. Performed the negation processing model in text sentiment analysis [41]. And their experiments analyzed the effects of both syntactic and descending negation keywords.

3 Proposed Model

The supervised and unsupervised machine learning techniques used on a previously classified dataset were considered almost accurate to predict the text data. The pre-classified datasets are generally domain-specific. Therefore, the machine learning model generates work only for a particular domain-specific text data. The text datasets are first converted into transitional models where text documents or corpus are represented as vectors. After then, the mutable representations are inserted into the machine learning algorithm. Our research found that Multinomial Naïve Bayes, Multivariate logistic regression, Max Entropy Random Forest, and Linear Support Vector Machines are popular algorithms for text classification.

Term Frequency (TF): Term Frequency and Inverse Document Frequency are two different words that are combined in the term frequency-inverse document frequency. To begin, the phrase "term frequency" refers to the number of times a term appears in a text document [42]. For example, the overall length of text documents might range from extremely short to extremely long, therefore each phrase may appear more frequently in large text documents than in tiny ones. To tackle this problem, the frequency of the times is calculated by dividing the existence of a term in a text document by the total number of terms in that text document. The term frequency of the word 'Alphabets' in the 'T1' text document is TF = 10/6000 = 0.0016 in another example.

Inverse document frequency (IDF): When computing the term frequency of a text document, the inverse document frequency is used. All text keywords are treated similarly by the algorithm approach. It makes no difference whether an English stop word like 'from' is incorrect. All text terms have distinct meanings. Let's say the stop word 'of' appears 3000 times in a text document, however it has no purpose or has very little meaning, which is useful for the inverse document frequency. In the text document, the inverted document assigns less weight to frequently used terms and more weight to fewer common words. For example, if we have ten receipts and five of them are text labelled 'technology,' the reverse document frequency is 4. Log e (10/5) = 0.301. IDF = log e (10/5) = 0.301.

Term Frequency - Inverse Document Frequency (TF-IDF): Inverse Document Frequency - Term Frequency indicates that in a text, more or higher occurrences of a text word in documents lead to a higher term frequency. In a text, documents with fewer words have a higher importance (Inverse

Document Frequency) for the text keyword sought in that document. The term frequency-inverse document frequency is simply the product of the terms frequency and inverse document frequency multiplied together (IDF). In the preceding sections, the calculations of term frequency and inverse document frequency are discussed. TF-IDF = 0.003 * 0.301 = 0.000903 is the formula for calculating term frequency-inverse document frequency.

Every text word is turned into a number, and the text document or corpus is represented as a vector. In the term frequency and inverse document frequency model, this number might be binary zero and one (0 and 1) or any actual integer. If a text word appears in a document, it receives a score of 1 for the text keyword, and if the word does not appear, it receives a score of 0 for the text keyword in a binary bag of words technique. As a result, the document or corpus vector is just a series of 1s and 0s. The document vector can be a list of any numbers calculated using the term frequency-inverse document frequency approach in the case of term frequency and inverse document frequency.

The goal of our tf-idf and n-gram model technique is to use the term frequency-inverse document frequency vectorization technique for text classification and the n-gram technique for word extraction. To accomplish predictive modelling, four different methods were chosen: Logistic Regression, Random Forest Classifier, Multinomial Naive Bayes, and Linear Support Vectorization. As detailed in the study, the dataset was randomly divided into 70 percent -30 percent training and testing portions.

4 Data Collection

The collection is based on the Consumer Complaint Database at catalog.data.gov [43], which contains complaints about a variety of consumers financial products and services. From a variety of sources, the consumer complaint database collects complaints regarding consumer financial products and services that are delivered to corporations for response. After the corporation responds and verifies a commercial relationship with the consumer, or after 15 to 20 days, whichever occurs first in a cycle, the customer complaints are published. The other valve was the subject of the complaints. The database is maintained on a daily basis. Some examples of categories in the data collection are a loan, a student loan, money services, and a savings account., etc. (see Fig. 1)

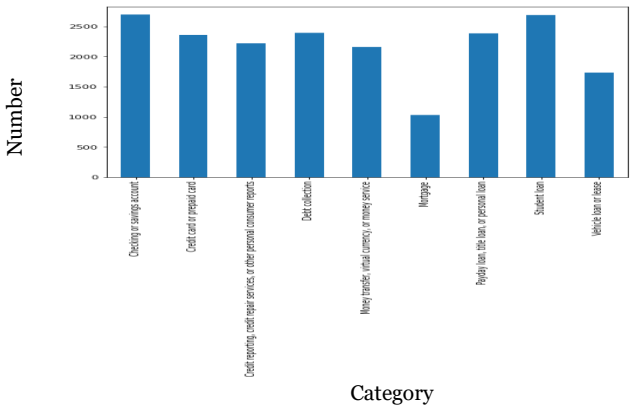

Fig. 1. The graph represents the different categories considered from the data set and its counts

Data Pre-Processing

Only two columns were used in the training data set: 'Product' and 'Consumer complaint narrative.'In the data set, the number of complaints by product is unbalanced. When looking at the data, the customer complaints are more inclined towards debt collection, student loans, mortgages, and other categories. When they run into problems like these while undertaking machine learning, they're going to have a hard time fixing them using normal algorithm strategies. Traditional algorithms are biassed towards the dominant class in the model and do not take into account the data distribution appropriately. Different classes are viewed as outliers in the worst-case scenario, and they are ignored or overlooked, resulting in bias in the machine learning model. Some applications, such as fraud detection or cancer prediction, may necessitate carefully configuring the model or artificially balancing the dataset by under- or over-sampling each class. Most classes, on the other hand, might be of significant relevance in our use case of learning an imbalanced data set. It's ideal to have high forecast accuracy for the majority of classes while retaining reasonable accuracy for the opposition. As a result, I'll leave it alone.The machine learning classifiers and learning algorithms cannot process text documents directly in their original form because most expect fixed-size numeric feature vectors rather than the raw text documents with variable length. Therefore, the texts are converted to a more manageable representation in the pre-processing step.

Because most machine learning classifiers and learning algorithms require fixed-size numeric feature vectors rather than raw text documents with variable length, they cannot analyse text documents in their original form. As a result, in the pre-processing step, the texts are changed to a more comprehensible format.

The bag-of-words model is the typical machine learning approach for extracting features from text sources. The paradigm in which a complaint expressing the presence (frequency) of text terms is taken into account for each text document, but the sequence in which they occur is ignored. We will generate a metric termed term frequency-inverse document frequency, shortened as tf-idf, for each keyword term in our dataset. For each consumer complaint narrative, we'll create a term frequency-inverse document frequency vector. And, during the pre-processing stage, the'sublinear' option is set to 'True' to use the logarithmic frequency form. A'min df' is the minimal number of text documents that must be present in a word in order for it to be saved. To ensure that all feature vectors have a euclidian norm of 1, the norm is set to a 'l2'. The 'ngram range' is set to one and two (1 and 2) to indicate that unigrams and bigrams approaches would be considered. The'stop words' variable is set to 'English' to remove all common pronouns ("a," "the,"...) from the text document, reducing the quantity of noisy features. The duplicate sentences from the text document were taken into account at the end of the pre-processing stage.

5 Implementation (Machine Learning Method)

The goal of this paper was to see if it was sufficient to regard text classification as a subset of topic-based text categorization or if new text categorization methods were required. I tried Naive Bayes classification, logistic regression, linear support vector machines, and random forest, as well as four other standard algorithms. These four methods have different working formulas, but each has proven to be effective in past text classification studies. Using the conventional bag-of-words architecture, build these machine learning methods on a consumer finance data set. Let f1,..., fm represent a predetermined set of m features that can appear in a text document, such as the word'stills' or the bigram'really stinks.' The number of times 'f1' appears in the document 'd' is n1(d) (2). The document vector then represents each text phrase 'd'.

$$d := (n1(d), n2(d), \ldots, nm(d)) \qquad (2)$$

Then text features are used to represent each consumer finance complaint storey, resulting in the term frequency-inverse document frequency score for various unigrams and bigrams methodologies. Also, the feature selection chi2 may be used to determine the most correlated words with each product and the sentences that go with it.

The 'customer complaint narrative' was converted into a vector of integers, and then supervised classifiers were trained using vector representations such as word frequency-inverse document frequency weighted vectors. They fall under respective text sentences after having these vector representations of the text in the document, which may train the supervised classifiers to train unseen finance 'customer complaint narrative' and predict the 'product.' After you've completed all of the data set transformations and have all of the features and labels, you can start training the classifiers. Once the model selection is complete, the next step is to experiment with several machine learning models, evaluate their accuracy, and identify the source of any potential model faults. Finally, the following four models will be benchmarked. Random Forest (RF), Logistic Regression (LR), (Multinomial) Naive Bayes (NB), Linear Support Vector Machine (LSVM), and Linear Support Vector Machine (LSVM) (RF).

Based on the classification with four prediction models achieved the following results.

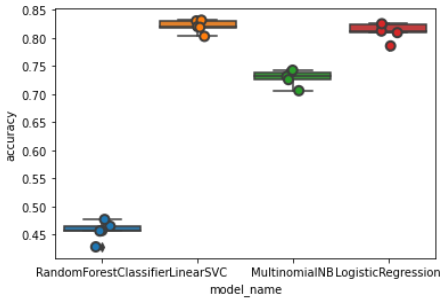

Fig. 2. The chart represents the using ensemble technique with the different algorithms representing each algorithm's accuracy value

The linear support vector classifier and logistic regression outperform the other two classifiers in this case (see Fig. 2). With a median accuracy of roughly 84 %, the linear support vector classifier (LinearSVC) has a little advantage.

Model Evaluation

We analysed the confusion matrix and presented the difference between predicted and real labels using our best model (LinearSVC).

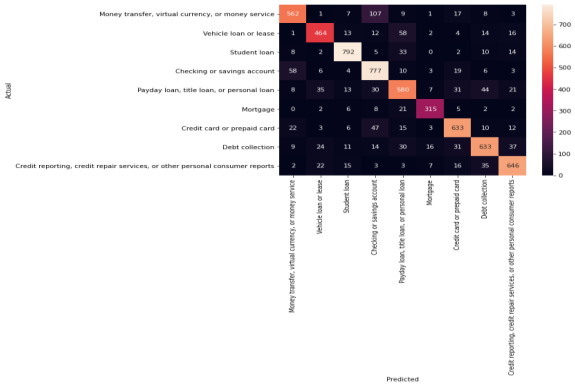

Fig. 3. The chart represents the ensemble technique with the different algorithms representing each accuracyvalue

The majority of predictions result in diagonal values (predicted label = actual label). In the category between manually categorised and machine learning classified based on the data set training in the model, there are multiple misclassifications (see Fig. 3). Some complaints have been misclassified, which means that classes from the data set cover more than one issue (for example, complaints concerning a credit card and credit reports). Then, using the chi-squared test, locate the most correlated terms with each group to achieve an accurate answer once more. They are consistent with our expectations after testing. Finally, print out each class's classification report.

Table 1. Model accuracy for each classification category

Category	Precision	Recall	F1-Score	Support
Money transfer, virtual currency	0.85	0.80	0.82	716
Vehicle loan	0.84	0.80	0.82	585
Student loan	0.92	0.92	0.92	867
Checking or savings account	0.78	0.89	0.83	888
Payday loan or personal loan	0.77	0.76	0.77	770
Mortgage	0.90	0.88	0.89	362
Credit card or prepaid card	0.85	0.85	0.85	752
Debt collection	0.84	0.80	0.82	806
Credit reporting or other personal consumer reports	0.87	0.87	0.87	750
accuracy			0.84	6489
macro avg	0.85	0.84	0.84	6489
weighted avg	0.84	0.84	0.84	6489

6 Conclusion

In this paper, we have conducted experiments on consumer data sets with four different algorithms that have demonstrated an effective and efficient method of text classification using the term frequency-inverse document frequency and n-gram model. The final model using the linear support

vector classifier algorithm shows an overall efficiency of 84% (see Table. 1). Student loan-related classification is the most accurate among other classes from the data set, having an f1-score of 92%. We have concluded that when the term frequency-inverse document frequency model and the combination with n-gram technique, then the performance of the classifier increases by a good percentage for the machine learning technique, and new sentences can predict with high accuracy without any human intervention or human help not required. And also, can be implemented data set using different models like word2vec and Bert to check the accuracy level with the current model for future analysis. And there are several directions to improve or develop the proposed work. In the future, the recommended model could be extended to the cross-domain aspect extraction technique so that the aspect terms technique accomplished from one domain can be used to analyze the aspect terms of some other discipline. This can be achieved by identifying a domain-independent feature set for token representation in natural language processing.

References

[1] Patel, D. et al. (2020). Implementation of Artificial Intelli-gence Techniques for Cancer Detection. *Augmented Humen Research*, 5(6):

[2] Patel, D., Shah, D. and Shah, M. (2020). The Intertwine of Brain and Body: A Quantitative Analysis on How Big Data Influences the System of Sports. *Annals of Data Science*, 7: 1–16.

[3] Patel, H. et al. (2020). Transforming petroleum downstream sector through big data: a holistic review. *Journal Petroıuml Exploration and Production Technology*, 10: 2601–2611.

[4] Panchiwala, S. and Shah, M. (2020). A Comprehensive Study on Critical Security Issues and Chal-lenges of the IoT World. *Journal of Data, Information and Management*, 2: 257–278.

[5] Hariri, R. H., Fredericks, E. M. and Bowers, K. M. (2019). Uncertainty in big data analytics: survey, op-portunities, and challenges. *Journal of Big Data*, 6: 44.

[6] Jaseena, K. U. and David, M. J. (2014). Issues, Challenges and Solutions: Big Data Mining. *Computer Science & Information Technology*, 4: 131-140.

[7] Ahir, K. et al. (2020). Application on Virtual Reality for Enhanced Ed-ucation Learning, Military Training and Sports. *Augmented Humen Research*, 5.

[8] Gandhi, M., Kamdar, J. and Shah, M. (2020). Pre-processing of Non-symmetrical Images for Edge De-tection. *Augmented Humen Research*, 5.

[9] Jani, K. et al. (2020). Machine learning in films: an approach to-wards automation in film censoring. *Journal of Data, Information and Management*, 2: 55–64.

[10] Jha, K. et al. (2019). A comprehensive review on automation in agricul-ture using artificial intelligence. *Artificial Intelligence in Agriculture*, 2: 1-12.

[11] Shah, D. et al. (2020). A Comprehensive Analysis Regarding Several Breakthroughs Based on Computer Intelligence Targeting Various Syndromes. *Augmented Humen Research*, 5.

[12] Shah, K. et al. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Humen Research*, 5.

[13] Kakkad, V., Patel, M. and Shah, M. (2019). Biometric authentication and image encryption for image security in cloud framework. Multiscale and Multidiscip. *Multiscale and Multidisciplinary Modeling, Experiments and Design*, 2: 233–248.

[14] Kundalia, K., Patel, Y. and Shah, M. (2020). Multi-label Movie Genre Detection from a Movie Poster Using Knowledge Transfer Learning. *Augmented Humen Research*, 5.

[15] Pandya, R. et al. (2020). Buildout of Methodology for Meticulous Diagnosis of K-Complex in EEG for Aiding the Detection of Alzheimer's by Artificial Intel-ligence. *Augmented Humen Research*, 5.

[16] Parekh, V., Shah, D. and Shah, M. (2020). Fatigue Detection Using Artificial Intelligence Framework. *Augmented Humen Research*, 5.

[17] David, J. M. and Balakrishnan, K. (2011). Prediction of Key Symptoms of Learning Disabilities in School-Age Children Using Rough Sets. *International Journal of Computer and Electrical Engineering*, 3.

[18]Talib, R. et al. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, 7.

[19]Fan, W. et al. (2006). Tapping the power of text mining. *Commun*ications of the *ACM*, 49: 76–82,

[20] Herranz, S., Palomo, J. and Cruz, M. (2018). Building an Educational Platform Using NLP: A Case Study in Teaching Finance. *Journal of Universal Computer Science*, 24: 1403-1423.

[21] Sukhadia, A. et al. (2020). Optimization of Smart Traffic Governance System Using Artificial Intelligence, *Augmented Humen Research*, 5.

[22] Pathan, M. et al. (2020). Artificial cognition for applications in smart ag-riculture: A comprehensive review. *Artificial Intelligence in Agriculture*, 4: 81-95.

[23] Lewis, C. and Young, S. (2019). Fad or future? Automated analysis of financial text and its implica-tions for corporate reporting. *Accounting and Business Research*, 49: 587-615.

[24] Kumar, B. S. and Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114: 128-147.

[25] Bafna, P., Pramod, D. and Vaidya, A. (2016). Document clustering: TF-IDF approach, In *the International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 61-66.

[26] Trstenjak, B., Mikac, S. and Donko, D. (2014). KNN with TF-IDF based Framework for Text Cate-gorization. *Procedia Engineering*, 69: 1356-1364.

[27] Gautam, J. (2013). An Integrated and Improved Approach to Terms Weighting in Text Classifica-tion. *International journal of Computer science*, 10:310-314.

[28] Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

[29] Mosteller, F. and Wallace, D. L. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, Springer-Verlag.

[30] Argamon-Engelson, S., Koppel, M. and Avneri, G. (1998). Style-based text categorization: What newspaper am I reading?. In *Proceeding of the AAAI Workshop on Text Categorization*, 1–4.

[31] Finn, A., Kushmerick, N. and Smyth, B. (2002) Genre classification and domain transfer for infor-mation filtering. In *Proceeding of the European Colloquium on Information Retrieval Research*, 353–362.

[32] Wiebe, J., Wilson, T. and Bell, M. (2001). Identifying Collocations for Recognizing Opinions. In *Proceeding of the ACL/EACL Workshop on Collocation*.

[33] Turney, D. P. and Littman, M. L. (2002). Unsupervised Learning of Semantic Orientation from a Hun-dred-Billion-Word Corpus. *Technical Report EGB-1094, National Research Council Canada*.

[34] Tong, R. M. (2001). An Operational System for Detecting and Tracking Opinions in On-Line Dis-cussion. In *Proceedings of SIGIR Workshop on Operational Text Classification*.

[35] Alison, H. and Pero, S. (2000). Fuzzy Typing for Document Management. In *ACL* 2000 *Companion Volume: Tutorial Abstracts and Demonstration Notes*, 26–27.

[36] Peter, T. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Computing Research Repository – CORR*, 417-424.

[37] Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 79-86.

[38] Mullen, T. and Collier, N. (2004). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 412-418.

[39] Michael, W. et al. (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 60-68.

[40] Farooq, U. et al. (2016). Negation Handling in Sentiment Analysis at Sentence Level. *Journal of Computers,* 12: 470-478.

[41]Zhang, Z. (2008). Weighing Stars: Aggregating Online Product Reviews for Intelligent E-commerce Applications. In *Intelligent Systems, IEEE*, 23:42-49.

[42] Hakim, A. A. et al. (2015). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse docu-ment frequency (TF-IDF) approach. In *6th International Conference on Information Technolo-gy and Electrical Engineering: Leveraging Research and Technology, (ICITEE)*.

[43] https://catalog.data.gov/dataset/consumer-complaint-database