# A Literature Review on Sentiment Analysis and its Foundational Technologies

Stavros Karmaniolos and Geoff Skinner
School of Electrical Engineering and Computing
The University of Newcastle
Newcastle, Australia
Geoff.Skinner@newcastle.edu.au

*Abstract*—**Sentiment Analysis is the computational treatment of opinion or sentiment expressed in a source of data. A cross-section of Natural Language Processing and Machine Learning, Sentiment Analysis applications deal with opinions as a tangible commodity often used to leverage big data for the sake of gaining a competitive advantage. To date it appears Sentiment Analysis has no formal frameworks or seminal texts that one can turn to for an accessible but comprehensive look into the field. This literature review based paper addresses this absence by providing a technical summary of common modern frameworks, the taxonomy of its parent fields and a brief analysis of related research dating back to the early 1990s.**

## I. Introduction

As unstructured data grows at an exponential rate in the information age, the difficulty found in deciphering this data automatically rises in proportion with the growth. In prior times, academia and enterprise alike used surveys to collect information to analyse potentially gainful information. These methods were slow, expensive and limited in almost every conceivable way when compared to current methods described in this paper. A modern approach to extracting knowledge from large datasets can be found in Sentiment Analysis [4]. Sentiment Analysis is the computational treatment of opinion, sentiment and subjectivity [1] and can be categorised as a shared field of both natural language processing (NLP) and machine learning (ML) which are both contained within the broadly defined field of artificial intelligence (AI) [17]. A recent surge of activity within this area can at least be partly attributed to the interest in dealing directly with opinions as a tangible commodity in areas including, but not limited to marketing, politics, decision support systems and finance [8,10,11]. Numerous studies [6,7] have shown that online consumer reviews have a significant impact on offline purchasing behaviour with 81% of American internet users stating that they have researched a product online, 20% do so daily and between 73% and 87% report that reviews had influenced their purchasing decisions. Consumers also are willing to pay from 20% to 99% more for five-star rated products than a four-star rated product [6].

The value proposition of a large amount of readily available data is apparent to researchers and industry alike [1,8,19] and this growth has lead to more effective frameworks being continuously developed. This paper will offer a look at the different classification methods used in understanding opinion from unstructured text. In particular, focusing on the most successful and widely used methods of processing and categorising text. Additionally, this paper sheds light on the background, history and applications of Sentiment Analysis.

## II. Background

The study of the polarity and semantic orientation of words has been explored extensively, particularly from the mid-to-late 1990s. Early work by Huettner and Subasic [32] and Hatzivassiloglou and McKeown [33] focussed on the study of the semantic orientation of words. The former developed a cognitive linguistic model for sentiments based on fuzzy logic [5]. The latter pair of researchers used a log-linear regression model in order to predict the orientation of two conjunctively joined adjectives [33].

In the field of machine learning, both [33] and [34] use similar rules to extract syntactic structures to assist in the identification of the orientation of specific cases for domain oriented semantic analysis [21]. Other early research performed by Guralnik and Karypis [35] as well as Han and Baker [36,37] and Zhong, Altun, Harrison [39] explored using scalable vector machines and k-means clustering to identify patterns and classify texts in a variety of domains including information security and bioinformatics.

Few works exist that focus on analysing sentiment using words without extracting representative features [21]. Words that project an opinion are classified and then analysed to decipher polarity in [41] whereas [42] identified the sentiment expressions and terms of each sentence. Nasukawa and Yi focussed on a lexical rule-based method of recognising emotion from text and founded the “Affect Analysis Model” which was applied in the 3D game world, “Second Life” [41].

More recent research focuses on extracting opinion by pairing data mining algorithms and feature selection methods [1-5,8,11-15]. Pang and Lee determined whether a review was positive or negative by classifying the data by overall sentiment rather than the topic. The findings show that machine learning techniques outperformed human-produced baselines when combining data mining and feature selection methods [12]. Bing and Lei focused on decision-making based on opinions found online through reviews, blogs and social networks [1]. In this study, the researchers used data mining algorithms in conjunction with feature selection methods to identify spam or fake reviews which may have an unwanted effect on the purchasing behaviour of consumers online and offline [1].

## III. Contemporary Technologies

In natural language, opinions can often be expressed in complex, subtle ways [2]. Simple text-categorisation such as n-gram or keyword identification are often ineffective and cumbersome when working with large, intricate data sets [13]. The escalated growth of platforms that facilitate user generated content presents an opportunity to gain deeper insight into consumer opinion about topics, brands or people [7]. Sentiment Analysis is often a lengthy process containing many steps including gathering and cleaning data, training models, training classifiers and extracting features [20,43], however, the entire process is typically abstracted into three main aspects: processing, classification and validation. While no formal framework currently exists [31], this section will cover differing techniques used by researchers and enterprises to analyse and gain competitive advantage from unstructured data. We will not cover the acquisition of large data sets in any depth however it is worth noting that data can be easily extracted from publicly available data sources like review websites, blogging sites, forums and social media websites [1,21].

### *A. Document Processing*

The first major element of Sentiment Analysis depicts the researcher cleaning data in preparation for later classification. Fanny, Muliono and Tanzil used a tokenisation processing method outlined in table I during a 2018 study that conducted multiple classification experiments on various news sources including ABC news, Fox News, New York Times and BBC news [16]. While this task can consist of many, or a few steps, we will outline three in this paper which are commonly found in most of the reviewed literature.

TABLE I
Parameters for Document Processing [16]

| **Parameter** | **Mode Uses** |
|---|---|
| Tokenisation | Non-letters mode |
| Filtering stopwords | English standard |
| Stemming | Lovins, porter, wordNet |

*1) Tokenisation:* A token can be defined as a categorised block of text in a sentence [17] corresponding to a categorised function it performs [2]. So long as it is a useful part of the text, the categorised block can be considered a token. Typically this means splitting on punctuation and spaces depending on the source of data. For example, Twitter users will commonly use words like 'don't', 'I'll' and 'she'd'. In this case, splitting on punctuation marks would not yield a desirable result [45].

*2) Stopword Filtering:* A stopword is any word which will not be useful in classification since no useful information can be derived from it [16]. While there is no standardised list of stopwords, most tools use a set that includes articles ('a', 'an', 'of', 'the') [45], auxiliary verbs ('is', 'are', 'was') and context-specific words dependent on the dataset [17].

*3) Stemming:* The third filtration that should take place is known as stemming [2,45]. It details the reduction of words to the root. An example of this can be observed with the words "cleans" and "cleaning" which both possess the same meaning as "clean" so the affixes are removed [17].

### *B. Classification*

An early issue in the field of Sentiment Analysis was whether it could be simply treated as a case of topic-based categorisation ($T_1$ = Positive Sentiment, $T_2$ = Negative Sentiment) or whether further development was required for proper categorisation [3,12]. By learning patterns, variables and features from previously labelled items, objects or events in the existing data, classification methods categorise new instances into the correct groups or classes [17]. Clustering, another data mining method process, may be used to determine class attribution of features through an unsupervised learning process [30] where the algorithm only considers the input variables [17,35,39]. This paper will not explore clustering or all of the ways sentiment can be classified, like those found in [38] and [40], but instead focus on three commonly used machine learning approaches to deal with unstructured text data in the way Sentiment Analysis requires.

TABLE II
Classification Methods [16]

| **Classifier** | **Type** |
|---|---|
| **SVM** | RBF |
| | Linear |
| | Sigmoid |
| **Naive Bayes** | Estimation Mode: Full |
| **k-NN** | Euclidean Distance |
| | Cosine Similarity |
| | Correlation Similarity |

*1) Support Vector Machines:* A powerful collection of linear models, Support Vector Machines (SVMs) are a useful technique for the classification of a variety of inputs [17,22,23]. In this field, SVMs are primarily used as a tool to classify the tone of communication as positive or negative, an important step in the overall analysis [1,12,13]. Since 1991, SVMs have acted as an efficient mode of pattern recognition and regression estimation [24], notably used in text and data mining applications since its inception [16]. The overall process, when applied to the topic of Sentiment Analysis, begins with finding a hyperplane, $\overrightarrow{w}$, in the training procedure that separates the document vectors into a class separate from others and the margin is as large as possible [12,13]. There lies a constrained optimisation problem; $c_j \in 1, -1$ is the correct class of document $d_j$, the solution

$$\overrightarrow{w} := \sum_j a_j c_j \overrightarrow{d}_j, a_j \geq 0, \tag{1}$$

where $a_j$ is found by solving dual optimisation problems, $\overrightarrow{d}_j$ such that $a_j$ is greater than zero is called *support vectors* in this instance since they are documents contributing to and falling on $\overrightarrow{w}$'s hyperplane [12,14].

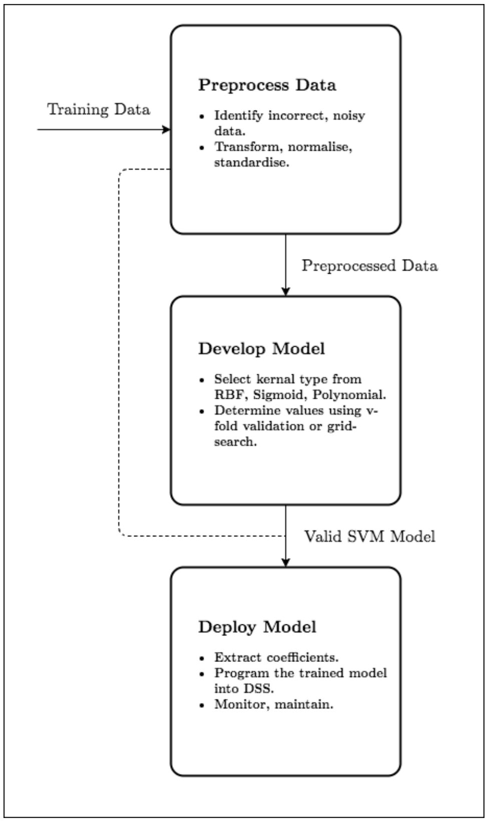

Training Data
Preprocess Data
• Identify incorrect, noisy data.
• Transform, normalise, standardise.
Preprocessed Data
Develop Model
• Select kernal type from RBF, Sigmoid, Polynomial.
• Determine values using v-fold validation or grid-search.
Valid SVM Model
Deploy Model
• Extract coefficients.
• Program the trained model into DSS.
• Monitor, maintain.

Fig. 1. Developing SVM Models [17]

While being easier to work with than artificial neural networks, users will often get undesirable results when using SVMs [15] often due to a lack of fundamental knowledge on domain intricacies [12,14]. In this instance, a process-based approach is sometimes employed. This consists of four major steps. Firstly, metricizing the data. This requires that each input instance is represented as a vector of real numbers [18]. Representing an $m$-class attribute where $m \geq 3$ using $m$ pseudo-binary-variables which, in practice, assume the value of 1 and others 0 based on the case class [17]. The next step is normalising the data. This is performed in order to omit values in high ranges dominating smaller range attributes. Large values may slow down the training process due to kernel values often depending on the inner aspects of vectors [11]. During this step, it is recommended that each attribute is normalised to the range $[-1, +1]$ or $[0, 1]$. After normalisation, the user must determine which kernel to use and assign kernel and penalty parameter(s) [22], $C$. The final stage is to deploy the model and integrate it into the decision support system once an ideal SVM prediction has been constructed [17].

*2) Naive Bayes:* The second approach is the Naive Bayes classifier. As the name suggests, this probabilistic classification method is based on Bayes' theorem classification algorithm [25]. This method is still used for text classification because of its speed and ease of implementation [16,28]. In Sentiment Analysis we adopt the base model of Bayes' theorem and expand upon it. By Bayes' rule,

$$P(a|b) = \frac{P(a)P(b|a)}{P(b)} \tag{2}$$

where $P(d)$ does not participate in selecting the class, $c*$ and to predict the given term $P(d|c)$, we decompose by assuming $f_i$, the predefined set of features that could appear, is conditionally independent given $d$'s class:

$$P_{NB}(c|d) := \frac{P(c)(\Pi_{i=1}^{m} P(f_i|c)^{n_i(d)})}{P(d)} \tag{3}$$

Pang, Lee and Vaithyanathan proposed a Naive Bayes implementation for the purpose of classifying online movie reviews [12] and found that this method performed well despite its conditional independence assumption often being ineffective in real-world scenarios [12,26-28]. In addition to this, Rennie, Shih, Teevan and Karger [27] state that assumptions made in Naive Bayes classification lead to high efficiency however adversely affect the quality of results. They create a modified, highly-efficient Naive Bayes algorithm, TWCNB, which can be seen in the table below comparing Multinomial Naive Bayes (MNB) to Transformed Weighted-Normalized Complement Naive Bayes (TWCNB) and the Support Vector Machine (SVM) over several data sets.

TABLE III
Comparing MNB, TWCNB and SVM Classifiers [27]

| Data Source | MNB | TWCNB | SVM |
|---|---|---|---|
| Industry Sector | 0.582 | 0.923 | 0.934 |
| 20 Newsgroups | 0.848 | 0.861 | 0.862 |
| Reuters (micro) | 0.739 | 0.844 | 0.877 |
| Reuters (macro) | 0.270 | 0.647 | 0.694 |

An example of a Naive Bayes implementation as a Java method can be seen in appendix A.

*3) k-NN:* Contrasting the aforementioned algorithms, the $k$-nearest neighbour algorithm is a simple, competitive prediction method used in data mining [17,46]. An instance-based prediction method for classification and regression-type prediction problems, $k$-NN algorithm can be defined as classified by a majority vote of its neighbours, with the feature assigned to a class most common among the $k$ nearest neighbours, where $k = \mathbb{Z}^+$ [17,48]. Examples of $k$-NN can be found used for spam mail detection [48], social media Sentiment Analysis [19,49] and food process decision-making [18] among other areas. Though no explicit training step is outlined in the typical $k$-NN process [47], the neighbours are initially taken from a set of correctly classified cases [3]. Represented in the figure below, if $k = 1$, then the object is assigned to the class of its nearest neighbour.

Figure 2 shows variables $(x, y)$, circles and squares representing known cases and the intercept position representing an instance of a new case. The assignment to a class depends on the closeness or proximity of the intercept to whichever is the nearest of the two known cases, circles or squares. If $k = 3$, the assignment can be made to a

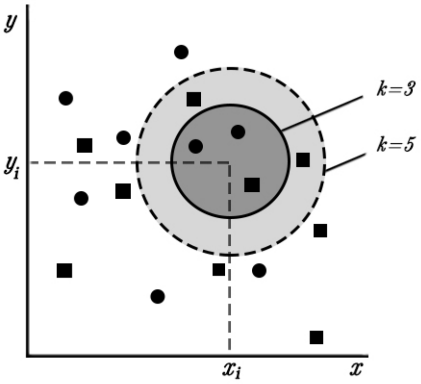


Fig. 2. Value of $k$ in $k$NN Algorithms [17]

circle whereas if $k = 5$, the value is a square [17].

There are two major decisions an analyst must make while using $k$-NN. Firstly, determining the similarity measure which is a mathematically calculable distance metric typically performed by three different functions. The Euclidean function calculates the linear distance between two points in a dimensional space [16,17] and is the most popular function. Manhattan is another commonly used algorithm and both of these are special cases of the Minkowski algorithm.

$$Euclidean : \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{4}$$

$$Manhattan : \sum_{i=1}^{k} |x_i - y_i| \tag{5}$$

$$Minkowski : \left( \sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{1/q} \tag{6}$$

The second decision is to decide the value of $k$. Depending on the data set, larger values of $k$ limits the space between class bounds and reduces noise [48]. An implementation of $k$-NN in Python can be found in appendix B.

## IV. Taxonomy

Artificial Intelligence (AI) is, in its simplest form, behaviour by a machine that, if performed by a human, would be called intelligent [17,51]. The concept can be traced back to 1956 at Dartmouth College in the U.S. conceived by four future Turing Award-winning researchers including Professors J. McCarthy, H. Simon, A. Newell and M. L. Minsky along with C. E. Shannon and N. Rochester [52]. Their original definition of AI was the ability of machines to understand, think, learn and emulate human intelligence. Since then, AI has grown significantly. Now considered a broad field consisting of highly specific sub-fields, AI has profoundly changed since the emergence of big data and the popularisation of the internet [53]. Applications of artificial intelligence systems can be found across most primary industries including health [53], supply chain management [4,18], politics [8] and social media [49]. AI integration with industrial structures continues to grow. IBM's 'Watson' system has been utilised by the health sector to increase the efficiency of screening patient records for historical instances of cancer treatments in order to provide a more effective leukaemia diagnosis method and provide schedules for therapy [53]. With roots in mathematics, computer science, engineering, philosophy among other fields, AI has branched off into specific application areas including those covered in this paper, Natural Language Processing and Machine Learning. The cross-section of these two fields is Sentiment Analysis [2,12,17].

### A. Natural Language Processing

Natural Language Processing (NLP) is a sub-field of Artificial Intelligence that studies how a computer can understand the natural human language, typically taking this language as input and converting into a formal representation suitable for computational manipulation [17]. Recently, the field has moved towards a data-driven methodology focussed on obtaining high-quality data in order to overcome the limitations of the traditionally employed rule-based system [54]. The typical data-driven NLP project life-cycle consists of steps not dissimilar to a data warehouse ETL (extract, transform, load) process [55]. Closing the gap between what a computer understands and sentiment expressed in a piece of text is the core goal of NLP research and it is not without challenges [17,54]. The following are some of the challenges that exist in NLP:

- Part-of-speech tagging;
- Text segmentation;
- Word sense disambiguation;
- Syntactic ambiguity;
- Imperfect or irregular input;
- Speech acts.

One area that has benefitted greatly from the advancements made in the computational analysis of language is politics. Automatic analysis of the opinions that citizens submit about regulations or policy proposals is an area where Sentiment Analysis is particularly effective [1,2,5]. In 2006, a team at Cornell proposed a model for electronic rulemaking or "eRulemaking" for short [8]. The identified issue was that each year federal regulatory agencies had to undergo a complex and expensive process called *notice and comment (N&C)* when issuing some of the approximately 4000 new rules proposed each year. This is the slowest rulemaking process and typically lasts between two to five years.

Initial results from the study were based off two issues and SVMs were used with five-fold cross-validation. The training used was issue-channeled comments, text of issues and manually identified relevant sections of the rules. The results yielded approximately 96% accuracy [8]. The broader impact of this project resulted in general-purpose natural language processing tools which are useful in any context requiring the management, organisation and analysis of large volumes of text [1,9]. The techniques that help agency rule makers can be used to design agency websites that help

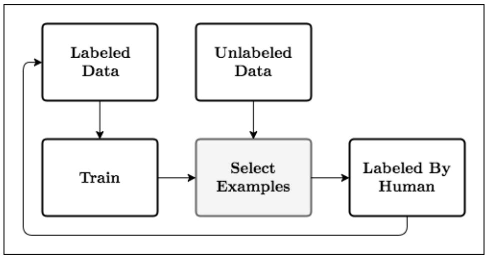


Fig. 3. Reducing Manual Annotation Costs: Active Learning [8]

citizens search, sort and selectively access materials relevant to the rule-making process. The project assisted in another study which determined, from the transcripts of U.S. congressional debates, whether speeches represent support of or opposition to proposed legislation [10].

### *B. Machine Learning*

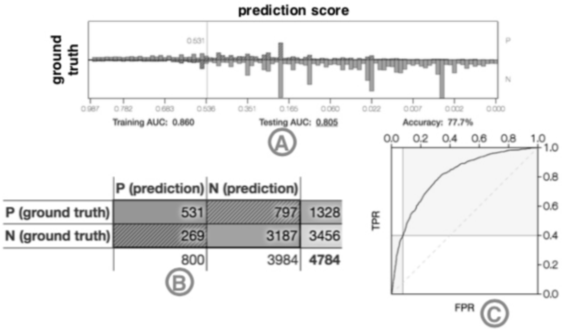
Machine Learning (ML) can be categorised into supervised or unsupervised learning [17,44]. In the context of this paper, ML is a set of algorithms concerned with treating Sentiment Analysis as a normal categorisation problem using syntactic and linguistic features [44,55]. Scalable Vector Machines, Naive Bayes and *k*-Nearest Neighbour algorithms are all popularly used in ML Sentiment Analysis [1] processes and are all supervised methods [44]. There are other methods including Maximum Entropy and Keyword-Based approaches [55] however these are not covered in this paper. While out-of-the-box ML solutions being readily available does have many benefits, the concept of a *black-box* has risen from the lack of understanding regarding the inner-workings of these systems by researchers, students and enterprise users alike [56]. Black-boxes can be found in commercial software like Weka, R, SVM Light and Matlab PRTools among others [55,56]. This means that often times researchers cannot explain the results of their work and often the result could be more effective if they possessed the knowledge about the classification method being used [56]. As ML models grow in complexity, the need for algorithm-agnostic approaches to explain the black-boxes becomes more valuable than ever.

J.W.H Krause [57] suggests the use of visual analytics to explain the black-box nature of Machine Learning. The proposed interface consists of three different panels, each corresponding to the different goals of the workflow proposed in the paper. The first panel is included below.

The future of AI will lead to autonomous machine learning that is both easily understood and more general in nature [53]. This progression will be partly attributed to the advancements made in examples like [56] and [57].

## V. Conclusion

Artificial Intelligence continues to pose some of the most interesting and challenging problems to modern computer scientists and its subsection, Sentiment Analysis is no different [16,54]. Contemporary technologies concerned with the computational understanding of sentiment have proven to be effective tools used by researchers, students and businesses [55] in understanding big data [4] however the pitfalls associated with using these *black-boxes* are present and pose a threat to the effectiveness and accuracy of the overall Sentiment Analysis process [56]. As researchers continue to propose new, innovative ways to mitigate these risks [57], the effects of these improvements will trickle-down into consumer products, bridging the divide between the sentiment expressed in unstructured text and the computational understanding associated with it.



| | P (prediction) | N (prediction) | |
|---|---|---|---|
| P (ground truth) | 531 | 797 | 1328 |
| N (ground truth) | 269 | 3187 | 3456 |
| | 800 | 3984 | **4784** |

Fig. 4. The Statistical Summary View for Machine Learning Black Boxes [57]

Examples of Sentiment Analysis were scattered throughout this paper to provide real-world instances of technology being used to gain competitive advantage and increase the overall effectiveness of business processes [1,2,4,11]. Products like Weka, SVM Light are readily available to anyone and Sentiment Analysis features are embedded into commercial products like Facebook Insights as well. In the future, we see the Sentiment Analysis embedding itself into nearly all businesses across all fields, specifically in marketing and advertising where understanding customer opinions and market sentiment is extremely valuable [2,7]. One other area of predicted growth is online forum moderation. Based on the findings in [38], we believe Sentiment Analysis tools can be used to combat online community toxicity and automate moderation of these forums in an efficient and effective manner.

## References

[1] L. Bing and Z. Lei, *Mining Text Data: A survey of Opinion Mining and Sentiment Analysis*. Boston, MA: Springer US, 2012.

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[3] B. Pang and L. Lee, *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*. Barcelona, Spain: Association for Computational Linguistics, 2004.

[4] P. Goel, A. Datta and S. Mannan, "Application of Big Data analytics in process safety and risk management," presented at 2017 IEEE International Conference on Big Data, Boston, MA, 2017.

[5] X. Bai, "Predicting Consumer Sentiments from Online Text," *Decision Support Systems*, vol. 50, no. 4, pp. 732-742, 2011.

[6] J. A. Horrigan, "Online shopping, Pew Internet & American Life Project Report, 2008.

[7] comScore, "Online consumer-generated reviews have significant impact on offline purchase behavior, Press Release, https://www.comscore.com/insights/press-releases/2007/11/online-consumer-reviews-impact-offline-purchasing-behavior?cs_cc=US, 2007.

[8] C. Cardie, C. Farina, T. Bruce, and E. Wagner, "Using natural language processing to improve eRulemaking, in Proceedings of Digital Government Research (dg.o), San Diego, CA, 2006.

[9] C. Sunstein, "Practically Binding: General policy statements and notice-and-comment rulemaking," *Administrative Law Review*, vol. 68, no. 3, pp. 491-516, 2016.

[10] M. Thomas, B. Pang, L. Lee, "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts," in Proceedings of EMNLP, Sydney, Australia, 2006.

[11] N. Li, D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection," *Decision Support Systems*, vol. 48, no. 00, pp. 354-368, 2010.

[12] B. Pang, L. Lee, "Thumbs Up? Sentiment Classification using Machine Learning Techniques," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79-86, 2002.

[13] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 412418, July 2004.

[14] T. Joachims, "Making large-scale SVM learning practical," Advances in Kernel Methods, Cambridge, MA, 1999

[15] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

[16] F. Fanny, Y. Muliono, F. Tanzil, "A Comparison of Text Classification Methods k-NN, Nave Bayes, and Support Vector Machine for News Classification," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 3, no. 2, pp. 157-160, 2018.

[17] R. Sharda, D. Delen, E. Turban, *Business Intelligence and Analytics,* 10th ed. London, GB, 2014.

[18] L.Peng, L. Wen, Y. Li, L. Qiang, D. Min, D. Yue. "Knowledge Acquisition Approach Based On Svm In An Online Aided Decision System For Food Processing Quality And Safety," *Cybernetics and Information Technologies*, vole. 14, no. 5, pp. 118-128, 2014.

[19] K. Denecke, W. Nejdl, "How valuable is medical social media data? Content analysis of the medical web," *Information Sciences*, vol. 179, no. 12, pp. 1870-1880, 2009.

[20] A. Bukhari, U. Qamar, "Critical Review of Sentiment Analysis Techniques," in Proceeding of the International Conference on Artificial Intelligence and Computer Science, 2014.

[21] M. Eirinaki, S. Pisal, J. Singh, "Feature-based opinion mining and ranking," *Journal of Computer and System Sciences*, vol. 78, no. 4, pp. 1175-1184, 2012.

[22] Z.H. Sun, Y.X. Sun, "Fuzzy support vector machine for regression estimation," *IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, no 8, pp. 3336-3341, 2003.

[23] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[24] J. C. Platt, B. Schlkopf, C. J. C. Burges, A. J. Smola, "Fast training of support vector machines using sequential minimal optimization" in Advances in Kernel Methods Support Vector Learning, MIT Press, pp. 185-208, 1998.

[25] M.G. Kendall, A. Stuart, K. Ord, "Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory," *Journal of the Royal Statistical Society Series D (The Statistician)*, vol. 84, no. 407, pp. 450-500, 1988.

[26] D. D. Lewis, "Representation and learning in information retrieval," vol. 7, 1992.

[27] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," in Proceedings of the 20th International Conference on Machine Learning, pp. 616-623, 2003.

[28] A. McCallum, K. Nigam, "A comparison of event models for naive bayes text classification," in AAAI-98 workshop on learning for text categorization, pp. 4148, 1998.

[29] V. Vryniotis, "DatumBox: Naive Bayes Classifier", GitHub, 2018. [Online]. Available: https://github.com/datumbox/NaiveBayesClassifier. [Accessed: 20-Sep-2018].

[30] P. Parashar and S. Sharma, "A Literature Review on Architecture, Classification Technique and Challenges of Sentiment Analysis", *International Journal of Engineering Research*, vol. 5, no. 5, 2016.

[31] J. Hollander, E. Graves, H. Renski, C. Foster-Karim and A. Wiley, "Urban Social Listening," 1st ed. Palgrave Macmillan, 2018.

[32] A. Huettner, P. Subasic, "Fuzzy typing for document management," Association for Computational Linguistics, *Tutorial Abstracts and Demonstration Notes*, pp. 26-27, 2000.

[33] V. Hatzivassiloglou, K.R.McKeown, "Predicting the semantic orientation of adjectives," in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp. 174-181, 1997.

[34] H. Kanayama, T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," in Proceedings of 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 355-363, 2006.

[35] V. Guralnik, G. Karypis, "A scalable algorithm for clustering protein sequences," Proc. Workshop Data Mining in Bioinformatics, pp. 73-80, 2001.

[36] K.F. Han, D. Baker, "Recurring local sequence motifs in proteins," *Journal of Molecular Biology*, vol. 251, no. 1, pp. 176-187, 1995.

[37] K.F. Han, D. Baker, "Global properties of the mapping between local amino acid sequence and local structure in proteins," in Proceedings of the National Academy of Sciences of the United States of America, pp. 5814-5818, 1996.

[38] J. Thompson, B. Leung, M. Blair and M. Taboada, "Sentiment analysis of player chat messaging in the video game StarCraft 2: Extending a lexicon-based model," *Knowledge-Based Systems*, vol. 137, pp. 149-162, 2017.

[39] W. Zhong, G. Altun, R. Harrison, P.C. Tai, Y. Pan, "Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property, *IEEE Transactions in NanoBioscience*, vol. 4, no. 3, pp. 255-265, 2005.

[40] S. Kim, E. Hovy, "Determining the sentiment of opinions," in Procedings of the 20th International Conference on Computational Linguistics, 2004.

[41] T. Nasukawa, J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in Proceedings of the 2nd International Conference on Knowledge Capture, pp. 70-77, 2003.

[42] W. Zhang, C. Yu, W. Meng, "Opinion retrieval from blogs," in Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management, pp. 831-840, 2007.

[43] V. Qian, "Step-by-Step Twitter Sentiment Analysis: Visualizing United Airlines PR Crisis", *iPullRank*, 2018. [Online]. Available: http://ipullrank.com/step-step-twitter-sentiment-analysis-visualizing-united-airlines-pr-crisis/. [Accessed: 14- Oct- 2018].

[44] S. Redmore, "Machine Learning vs. Natural Language Processing," *Lexalytics*, 2018. [Online]. Available: https://www.lexalytics.com/lexablog/machine-learning-vs-natural-language-processing-part-1. [Accessed: 14- Oct- 2018].

[45] A. Pak, P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in: *LREc*, vol. 10, 2010, pp. 1320-1326.

[46] P. Devijver, J. Kittler, *Pattern Recognition Theory and Applications,* Berlin, Heidelberg: Springer, 1987.

[47] S. Tan, "An effective refinement strategy for KNN text classifier," Expert Syst. Appl., vol. 30, no. 2, pp. 290298, 2006.

[48] D. Sharma, "Experimental Analysis of KNN with Naive Bayes, SVM and Naive Bayes Algorithms for Spam Mail Detection," vol. 8491, no. 4, pp. 225228, 2016.

[49] C. Bliss, I. Kloumann, K. Harris, C. Danforth and P. Dodds, "Twitter reciprocal reply networks exhibit assortativity with respect to happiness," *Journal of Computational Science*, vol. 3, no. 5, pp. 388-397, 2012.

[50] R. Saxena, "DataAspirant: K-Nearest Neighbor Algorithm Implementation in Python", DataAspirant, 2016. [Online]. Available: http://dataaspirant.com/2016/12/27/k-nearest-neighbor-algorithm-implementaion-python-scratch/. [Accessed: 27-Oct-2018].

[51] S. Tanimoto, *The elements of artificial intelligence: an introduction using LISP*, Rockville, MD: Computer Science Press, 1987.

[52] D. Crevier, *AI: the tumultuous history of the search for artificial intelligence*, New York, NY: Basic Books, Inc., 1993.

[53] Y. Pan, "Heading toward Artificial Intelligence 2.0", *Engineering*, vol. 2, no. 4, pp. 409-413, 2016.

[54] I. Zeroual, A. Lakhouaja, "Data science in light of natural language processing: An overview," in The First International Conference On Intelligent Computing in Data Sciences, 2018.

[55] R. Kimball, M.Ross, W. Thornwaite, J. Mundy, B. Becker, *The Data Warehouse Lifecycle Toolkit*, Hoboken: John Wiley & Sons, 2011.

[56] A. Rocha, J. Papa and L. Meira, "How Far Do We Get Using Machine-Learning Black-Boxes?," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 2, pp. 1-25, 2012.

[57] J. Krause, "Using Visual Analytics to Explain Black-Box Machine Learning," Ph.D Dissertation, New York University, 2018.