Natural language understanding challenges for sentiment analysis tasks and deep learning solutions

Radha Guha[1], Tole Sutikno[2]
[1]Department of Computer Science and Engineering, Faculty of Computer Sciences and Engineering, SRM University, AP, India
[2]Department of Electrical Engineering, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Article Info

Article history:

Received Feb 18, 2022
Revised Jul 29, 2022
Accepted Aug 16, 2022

Keywords:

Deep learning
LSTM RNN
Natural language understanding
Opinion mining
Sentiment analysis
Text classification

ABSTRACT

When it comes to purchasing a product or attending an event, most people want to know what others think about it first. To construct a recommendation system, a user's likeness of a product can be measured numerically, such as a five-star rating or a binary like or dislike rating. If you don't have a numerical rating system, the product review text can still be used to make recommendations. Natural language comprehension is a branch of computer science that aims to make machines capable of natural language understanding (NLU). Negative, neutral, or positive sentiment analysis (SA) or opinion mining (OM) is an algorithmic method for automatically determining the polarity of comments and reviews based on their content. Emotional intelligence relies on text categorization to work. In the age of big data, there are countless ways to use sentiment analysis, yet SA remains a challenge. As a result of its enormous importance, sentiment analysis is a hotly debated topic in the commercial world as well as academic circles. When it comes to sentiment analysis tasks and text categorization, classical machine learning and newer deep learning algorithms are at the cutting edge of current technology.

Corresponding Author:

Radha Guha
Department of Computer Science and Engineering, Faculty of Computer Sciences and Engineering
SRM University AP
Mangalgiri, Guntur, India
Email: radha.g@srmap.edu.in

1. INTRODUCTION

To understand why autonomous sentiment analysis (SA) [1]–[6] has become so crucial, we must first grasp the context of today's big data era. The internet is now used for a variety of purposes, including e-learning, e-commerce, e-health, and e-entertainment. This life-changing experience of having access to novel digital media applications from anywhere at any time has been made possible by the researchers' relentless efforts to make the internet, or the world wide web (www), more intelligent, connected, and data-driven semantic web by incorporating machine-based data understanding. The read-only static web has changed to a more interactive and collaborative read-and-write social web. Web 3.0 will soon be released, which will boost the web's intelligence by allowing web-based applications to automatically exchange data and make accurate decisions. Twitter was started in 2006 in San Francisco as a microblogging site during the Web 2.0 period. Twitter is a service that allows users to publish short messages to the public in a variety of languages all around the world. Businesses may communicate with customers in real time via Twitter. Today, almost 500 million tweets are published every day, for a total of 200 billion tweets every year. Other social media platforms include Facebook, Instagram, LinkedIn, Flickr, and Blogger, where individuals can openly share their views on any topic.

Unstructured text data from online review sites, personal blogs, and social media platforms transmit a wealth of business insight that can be profitably harvested. A user can identify which characteristics of a product he likes, and which features he does not like in a tweet. Other people consider online reviews when making their own purchasing selections. Because the brand image of a product is dependent on online reviews, social media posts are a source of concern for a company. Sentiment analysis or opinion mining (OM) [4], [7]–[10] of massive amounts of text data on the web can have numerous businesses uses. Customers' emails and voice messages are routinely read by the customer service team to resolve concerns as quickly as feasible. If an email could be automatically classified topically, the responsible department might be notified to address the issue in the email as soon as possible. For negative sentiment emails, the damage control team can take immediate steps to soothe unsatisfied or irate customers. Many firms' customer relationship management (CRM) systems have begun to integrate sentiment analysis tools to automate the study of product reviews posted on social media. Other people's perspectives are important not only on an individual level, but also on an organizational level.

Aside from product and service ratings, users use social networking sites to share their thoughts on political agendas and government mandates. It is possible to detect users' or consumers' moods like love, happiness, wrath, or hatred through social media monitoring. A corporation or government can take curative activities on time by using affective computing. It can forecast whether a political leader will win or lose an election. In another case, if sedition speech is identified in an online forum such as Twitter, the user's account will be deactivated. Fake news on social media, on the other hand, can be halted because it spreads quickly and can have negative consequences for a firm and its brand reputation. It goes without saying that in the big data era, data mining, particularly text mining from external sources of data, provides enormous opportunities for business improvement. It is normal for a human to understand the sentiment represented in writing and act properly. But what can a machine do to help with the process automation? The answer is "data mining," which transforms raw data into information, knowledge, and wisdom for a business, as seen in Figure 1. Social media data can be mined for business intelligence to monitor items and brand images to boost corporate income. Figure 2 depicts the interaction cycle between business intelligence and data mining. Data mining generates intelligence from business data, which is subsequently delivered back into the firm. Machines must be able to grasp natural language to extract commercial knowledge from social media data. This is also true for text sentiment analysis [11].

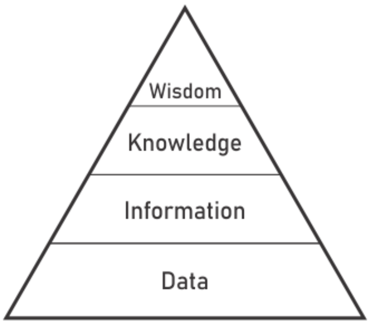

Figure 1. Data, information, knowledge, and wisdom (DIKW) pyramid

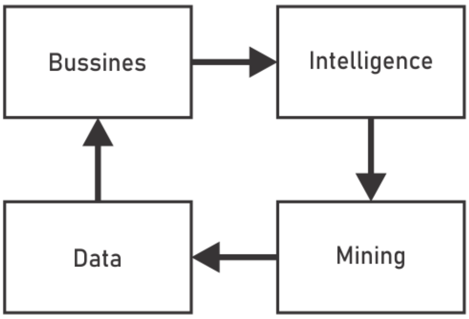

Figure 2. Business intelligence and data mining interaction

Natural language understanding (NLU) and sentiment analysis are also essential for machines to be called strongly artificially intelligent. With NLU to some degree, Siri, like the virtual assistant clubbed in the iPhone, is working as a conversational agent to fetch information from the internet as per the user's request. True NLU means understanding the semantic meanings and intent of the user's query rather than being just a keyword search. As there is a high demand for such applications, Google Assistant, Microsoft Cortana, and Amazon Alexa, like virtual assistants, have become very popular these days and have billions of customers. If a conversational chat bot can understand a user's sentiment, then it can answer more appropriately and will sound less robotic. With good NLU, including sentiment analysis, strong artificial intelligence (AI) products and services can be designed in the future.

In many great applications of NLU like machine translation, text summarization, and question answering sentiment analysis of text is a common sub-task. With SA, companies can automate customer preference sensing. From sentiment analysis of movie reviews, it becomes evident whether a movie is a box office success or a failure. Twitter sentiment analysis can help to predict stock market DOW Jones index movement well in advance by understanding users' moods. Airline tweets are very helpful for the airlines to take curative actions. There are many market analytics firms that do just that social media monitoring and sell the results to the respective company so that they can improve their brand image.

The SA tool can be used as part of a recommendation system [12], [13] designed to recommend the product to a user as per their taste expressed not only in quantitative rating but also in review text. If a product feature has received many negative comments, it will not be recommended to a user, or it can be sent back to the shop for redesign. Technology giants like IBM, Weka, SAS, SPSS, and Oracle have built sentiment analysis tools for paid use. But they are general-purpose tools and not fine-tuned for specific domains and do not tackle all challenges for better sentiment analysis. These commercial off-the-shelf (COTS) tools may fail when the sentence size is large enough to include natural language complexities such as negation, disjunction, anaphora, ambiguity, and sarcasm [14]–[19]. Even though the advancement of machine learning with deep learning neural nets is gradually improving NLU and SA accuracy, it is still far from a solved problem today.

For all the above-mentioned reasons, sentiment analysis is a popular research topic both in the business world and in academia to solve the open challenges of SA and to increase its accuracy. There are several supervised text classification algorithms like Nave Bayes (NB), decision tree, K-nearest neighbors (KNN), maximum entropy (ME), and support vector machine (SVM) that have been used for sentiment analysis also. Recently, deep learning techniques are also being used for sentiment analysis tasks. There are only a limited number of research papers which present all the challenges of sentiment analysis and then compare all the different algorithms and report their accuracy scores. Research papers published before 2016 reported sentiment analysis accuracy in the range of 60–70% on various social media datasets with different classification algorithms. More recent papers [20]–[22] published after 2016 report SA accuracy in the range of 70-90% with more complex deep learning neural network architectures like BERT and RoBERTA transformers based on transfer learning. Wide variations in performance are due to different dataset characteristics, their preprocessing, and neural network architecture.

This paper explores methodologies to improve the accuracy of automatic sentiment analysis of unstructured text documents and compares traditional and modern deep learning algorithms through experiments. Instead of relying on the claims of a limited number of papers and coming to a valid conclusion, we wanted to compare all the algorithms in this research paper. This paper verifies that deep learning solutions are better than traditional machine learning algorithms for sentiment analysis tasks. The paper's contribution is to present detailed challenges of the sentiment analysis task as well as experimental result comparisons of several traditional machine learning algorithms such as NB, logical regression (LR), SVM, and modern deep learning with basic MLP, CNN, and LSTM-RNN architecture. The rest of the paper is organized as follows. Section 2 discusses the details of subtasks to tackle the challenges of sentiment analysis. Section 3 describes the datasets and data exploration statistics. Section 4 compares results the experiments performed with traditional machine learning and modern deep learning techniques and discusses the outcome. Section 5 concludes the work done with future directions for this work.

2. CHALLENGES OF SENTIMENT ANALYSIS: RESEARCH METHODOLOGY

Sentiment analysis is one of the basic sub-tasks of NLU. SA, in turn, has numerous subtasks such as microtext preprocessing, subjectivity detection, named entity recognition (NER), aspect extraction, anaphora resolution, negation detection, and sarcasm detection [14]–[16], [19]. NLU and SA can be performed at three levels. First, at an individual word level, to understand word meaning and sentiments. Second, at sentence level, which is a composition of words, needs both syntactic and semantic analysis of the sentence. Third, at context level, which is a small paragraph of sentences of a review, a dialog, or a discourse. In a text, both subjective and objective sentences are present. Objective sentences are facts about something and are always true and neutral in sentiment. Subjective sentences are people's opinions about something like a product or event whose truth value has not been validated yet and conveys either positive or negative sentiment. First, subjective sentences are filtered out of objective sentences, and this is called subjectivity detection.

Sentiment analysis means emotion recognition and polarity classification of subjective sentences. The sentiment analysis task will take a piece of text as input and generate a scalar value as an output. This output can be -1, meaning negative sentiment, or +1, meaning positive sentiment. Or it can range from zero to five, meaning very negative, negative, neutral, positive, and very positive. After detecting the polarity of subjective sentences, affective computing will attach a label like "love," "happiness," "hate," "anger, and "deception" to the subjective sentences. Machine learning can happen in a supervised or unsupervised way. In supervised learning, already labeled training data of positive or negative sentiment reviews needs to be there. In unsupervised learning, learning happens in the absence of a labeled dataset.

Here is an example of a product review by a user. “Yesterday I bought an iPhone. It is great. Its touch screen is so cool. The voice quality is very clear. But my mother is angry with me as it cost me a fortune and she want me to return it". In this review, the first sentence is an objective sentence and has no opinion attached to it. The second, third, fourth, and fifth sentences are all opinion sentences, with the first three being positive and the last being negative. Overall, it is a mixed review. The aspect of the first positive opinion is the iPhone. The aspect of the second positive sentence is the touch screen. An aspect of the third positive sentence is the voice quality. An aspect of the last negative opinion sentence is the iPhone’s cost. The terms that are positive sentiment words are "great", "cool", and "clear". The negative sentiment phrase is "cost me a fortune" and the negative term is "return".

The challenge of SA arises as natural language like English is often ambiguous with the problem of synonymy and polysemy. In SA, word sense disambiguation is performed by a part of speech (POS) tagging algorithm. In English language sentences, words are tagged mainly into eight different categories, like article, noun, pronoun, verb, adverb, adjective, conjunction, and preposition. In Penn Tree Bank, a more detailed set of 45 categories is used. The POS tagging algorithm is the lowest level of syntactic analysis by word sequence labeling as per English grammar and will label the most probable category for a word in the context of the whole sentence. Only after knowing whether a word has been used as a noun or a verb, can its true meaning be deciphered. Verbs like "love," "hate," and "dislike" are sentimental words. Many adjectives have high sentiment value. Like in the sentence "It was an extremely embarrassing performance," the word "embarrassing" is an adjective with negative sentiment and "extremely" is an adverb that intensifies the negative sentiment score of the adjective. Most of the time, adjectives conjoined with ‘and’ have the same polarity, but those conjoined with ‘but’ have different polarities, like in the examples of ‘fair and legitimate’ and ‘enchanting but irrational’.

Over time, a few opinion or sentiment lexicons like MPQA, SentiWordNet, and LIWC have been compiled, which list positive and negative sentiment words. These lexicons not only specify positive vs. negative categories, but also specify strong vs. weak, active vs. passive, overstated vs. understated, or pain, pleasure, virtue, motivation like orientation of word opinions. For sentiment analysis of a document, Python NLTK’s opinion lexicon can be used. Examples of positive sentiment words are "love", "wonderful", "beautiful", "amazing", and "delicious", and negative sentiment words are "bad", "hate", "horrible", "terrible", and "sucks". Unsupervised learning can use these lexicons and many rules for word-word co-occurrence to score new words. To classify a document the computed sentiment score is: $\frac{\#positive\ terms'score-\#negative\ terms'score}{\#total\ words}$. But without the whole sentence words context, anaphora resolution and negation detection, unsupervised learning will not be very accurate. In absence of labeled example dataset, lexicons help in baseline sentiment classification.

Unstructured text preprocessing is called normalization of data and is a necessary step before the text data is numerically represented for machine learning algorithms. Preprocessing efficiency will determine subsequent sentiment classification accuracy. For natural language, preprocessing includes lower case, punctuation removal, stop word removal, stemming, and lemmatization for reducing the dimensionality of the data. Stop words are the most frequently used words in a corpus and are usually removed in the preprocessing step. But for SA, it is better to handpick stop words rather than blindly apply the stop word dictionary available in software packages like NLTK, Spacy, and SkLearn. For example, "not", "no", "never", "cannot", "shouldn’t", "without", and "may" are regarded as stop words and are often removed from a text for topic classification of a corpus. But for sentiment analysis, negation reverses the meaning of a positive sentiment phrase or sentence and should be preserved. For example, "not good" means negative sentiment, but "not bad" means positive sentiment. Negation and its scope detection is an important subtask for SA. For SA, the stop word list can contain only prepositions and determiners. For sentiment analysis, punctuation marks like "interrogation" and "exclamation" also hold some meaning and should not be removed.

Word tokenization and subsequent POS tagging are more challenging for Twitter like social media microtext. Microtexts are short texts where the message limit is 280 characters. All social media sites like Twitter, Instagram, and Facebook use many out of standard vocabulary terms (e.g., "HBD" for "Happy Birthday" or "LOL" for "laugh out loud", "b4" for "before" and "U" for "you"). The purpose of the shortening of the words is to get an increased speed of writing and to circumvent the length limit of tweets. Other times, a user stretches a word (e.g., 'goood’ for ‘good’ and 'sooo much’ for 'so much’) for exaggerating emotions. Social media users also use a lot of image emoticons or emojis, and they need to be mapped to respective emotion words. Spelling corrections and mapping of vocabulary terms to standard English terms is called normalization and is an added challenge for microtext sentiment analysis. Twitter text may also have special characters and extra blank spaces that need to be removed to clean up the text. Use of ‘@’ and ‘#’ hashtags is very common in tweets to indicate email addresses and trending tweet topics and can be cleaned as they are not meaningful for sentiment analysis. Microtext preprocessing is challenging to improve the accuracy of sentiment analysis models.

NER is the task of categorizing a word in natural text as being the name of a person, a place, an organization, an event, a date, or a price, or a specific feature such as the weight and color of a product. Sentiment words are about some aspects or named entities of a product or event. Aspect extraction means identifying the target of an opinion in a subjective sentence. Finding those aspect words is called NER. A product may have many features or aspects, and the review may be positive about one aspect but negative about another. Suppose a review is about a phone or a camera. A phone or camera has many aspects like brand, price, size, weight, lens, resolution, and battery life. A machine learning algorithm is applied to select aspect words and corresponding sentiment words or phrases from the documents.

Anaphora resolution and sarcasm detection are very difficult tasks in sentiment analysis. Anaphora means reference to an item mentioned earlier in a sentence. For example, the sentence: "The trophy does not fit in the suitcase because it is too small. Here, the word "it" refers to the suitcase. However, if the sentence is changed to "The trophy does not fit in the suitcase because it is too big," then the word "it" will refer to the trophy. In sarcasm, people use positive words to mean something negative. For example, the sentence "That is just exactly what I needed today!!!" This kind of anaphora resolution and sarcasm detection is challenging for traditional machine learning algorithms. Modern deep learning models do better in this regard when trained with a huge number of training examples. A deep learning network finds more meaningful non-linear input patterns to output mapping for sentiment classification.

Sentiment analysis involves several difficult subtasks that must be done one by one. Figure 3 shows the SA system architecture. After downloading Tweeter product reviews, they're tokenized, weighted, and saved in the form of sparse term-document matrix (TDM). Here each word is a feature in the term-document matrix (TDM) and all review documents have the same size numeric feature vectors which is equal to the corpus vocabulary size. And each feature vector length equals the number of documents in the corpus. This TDM model is a bag of words (BOW) model where word ordering in a document is not preserved. Term weighting is often a term's TF-IDF score. Sentiment classification is harder than topic classification, so term weighting can be varied and tested. For SA, presence (1) or absence (0) of a word, word-to-word coherence score, or pointwise mutual information (PMI) score are different choices employed as term weights. POS tagging, NER, anaphora detection, and negation handling can improve word weighting and feature selection for nouns, adjectives, and adverbs in a term document matrix (TDM) representation of data.

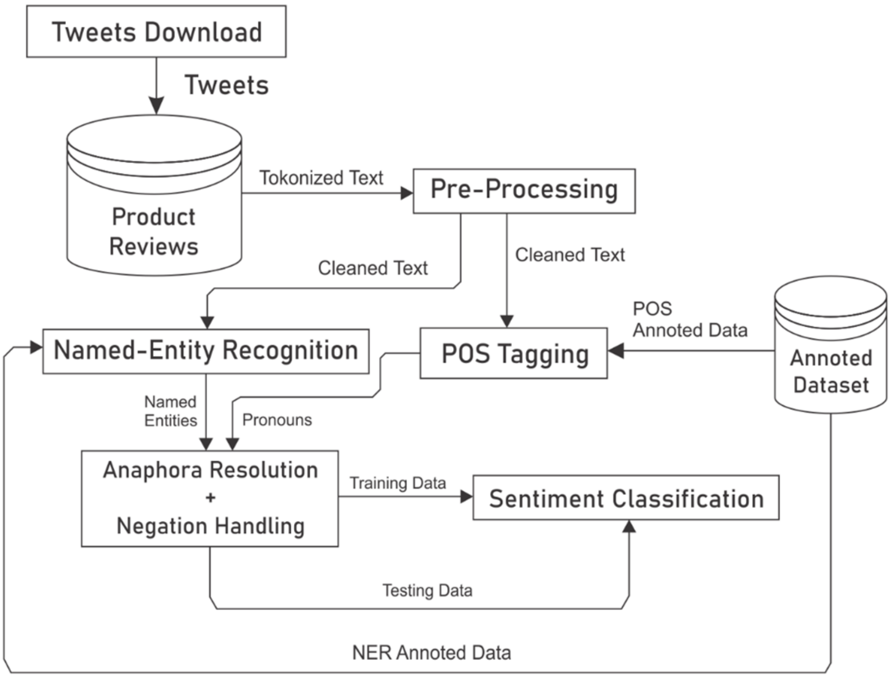

Figure 3. Sentiment analysis system architecture

The dimensionality reduction [23]–[29] of high dimensional sparse matrices, such as TDM, has progressed from latent semantic analysis (LSA) (1990) to latent dirichlet allocation (LDA) (2003) to principal component analysis (PCA) and non-negative matrix factorization (NNMF)-like algorithms. All of these approaches together represent a document vector within a much-reduced k-dimensional latent topic in the corpus. In this lower dimension representation of documents, the document-to-document similarity measured using a method comparable to the cosine distance yields more accurate results. For similarity computations of word vectors or document vectors, there are various notions of distance, like Euclidean, cosine, Jaccard, and Kullback-Leibler (KL) divergence. As there are so many different choices in every step, there is no other way than to experiment which combination of steps gives a more accurate similarity score.

But the latest word embedding schemes of Google's Word2Vec (2013) and Stanford's GloVe (2014) do distributional word representation in a much-reduced dimension and eliminate separate token reweighting and dimensionality reduction steps. This dense and reduced-dimensional word embeddings outperform sparse vectors in modern deep learning models [30]–[33]. The Word2Vec and GloVe models capture word context better and reduce the feature vector size to less than a 300-dimensional dense vector. Publicly available pretrained word embeddings trained on large internet datasets capture the semantic and syntactic meaning of words. This word embedding technique determines word to word relationships better and helps in more accurate word sense disambiguation, NER, parts of speech tagging, and sentiment similarity scoring of words. This new word embedding technique helps to detect sentiment conveyed in a text better. With this reduced dimension feature representation, the deep learning neural network produces better sentiment classification accuracy.

In 2015, deep learning began to show signs of improving its accuracy in areas such as computer vision, speech recognition, and natural language processing. Deep learning is particularly expensive to calculate since it consists of many hidden layers and a huge number of nonlinear neurons. However, modern computers with multi-core CPUs and GPUs have more computational capacity than they had in the past, which makes parallel processing conceivable. During the same period, the accessibility of large amounts of data throughout the course of the previous ten years has made deep learning feasible. Research on AI that makes use of deep learning has seen a meteoric rise in popularity over the past several years. Deep learning algorithms are more accurate than traditional machine learning algorithms like NB, maximum entropy, and SVM. These deep learning algorithms include numerous hidden layers and can pick up simple to complicated hierarchical patterns in multiple hidden layers. Experiments were carried out with a word embedding layer on the input side of the MLP feedforward architecture, the CNN architecture, and the long short-term memory (LSTM) recurrent neural network (RNN) architecture for the purposes of this article.

In the process of deep learning, a neural network is provided with an input sequence that is of a predetermined length. If a given document is smaller than this predetermined size, zeros are inserted to the beginning and end of it, but it is truncated if it is larger than the input form. Because the word order is maintained in this way, the neural network can learn more effectively. Python-Keras is a new deep learning framework that was created on top of the low-level TensorFlow ML platform to make the creation of deep learning models more user-friendly. Keras is versatile enough to run on a CPU, GPU, or TPU for the purpose of accelerating the execution of datasets. TensorFlow will search for computers that have an NVIDIA GPU and CUDA computing capabilities whenever there is a possibility that a program will allow for parallel processing.

In this study, we used fundamental architectures of deep learning to test how well they performed in comparison to traditional NB, LR, and SVM. In 2016, a comparison evaluation of different algorithms [30] indicated that distinct datasets and algorithms had an accuracy of between 60 and 70%. In 2017, improvements were made to applications that required NLU, such as question answering and automatic text summarization. These improvements were made possible by attention-based deep learning technology. Comparative research conducted in 2021 using deep learning architectures to test datasets reveals a variance of between 80 and 90%. Our study also validates superiority of deep learning over classical algorithms.

3. DATASET

The dataset of IMDB movie reviews that is freely available to the public and the dataset of Tweeter comments regarding airlines that is available on Kaggle are both utilized in this paper. Both the datasets were labeled for supervised machine learning. In our experiment our corpus is a collection of documents D={d1, d2, …, dn}. Each document *d* belongs to a class C={c1=positive, c2=negative, c3=neutral}. The dataset of IMDB reviews contains 49582 movie comments, roughly equally split between reviews with positive and negative polarity attitudes. The second dataset includes 14604 tweets on airlines, all of which fall into one of three categories: positive, negative, or neutral. When it comes to airline tweets, there is a significantly larger number of tweets with a negative mood than tweets with a positive or neutral sentiment. By tallying all the votes, we can determine which movies or airlines are the best and which are the worst, as well as which expressions are used to indicate positive and negative views. In Table 1, you will find a description of two dataset statistics that will assist you in better comprehending the results of the experiment. As shown in Table 1, the average number of terms used in individual movie reviews is significantly higher than that used in tweets about airlines, which may be one factor that contributed to the inferior accuracy of the IMDB dataset. A word cloud map displaying phrases indicating a positive reaction to a movie is presented in Figure 4.

Table 1. Dataset statistics

Dataset	Positive review	Negative review	Neutral review	#Documents or reviews	#Terms	#Unique terms	#Average terms per document
IMDB movie reviews	24,884	24,698	0	49,582	2,56,838	2,21,053	121
Airline tweets	2,354	9,159	3,091	14,604	30,165	26,910	17

Figure 4. Word cloud map with positive sentiments documents

4. RESULTS AND DISCUSSIONS

Following the execution of several preprocessing procedures, token reweighting and feature selection, the documents are then numerically represented in a TDM. In this article, the TDM is used as the input for three classic machine learning algorithms. These algorithms are known as NB, LR, and SVM. As part of the sentiment analysis process, our final task is to classify any newly published tweets or movie reviews which is not preclassified, by the trained model as expressing either a positive or negative sentiment.

Experiments are performed on Python’s Jupyter notebook. NLTK is a versatile Python library for NLU and natural language processing tasks. NLTK also provides a movie review dataset to experiment with. First the collected data is split in 80:20 train-test ratio. After building the model with train dataset it is evaluated with test dataset.

Precision, recall, accuracy, and F-score are the metrics that are used to evaluate model performance. Because the IMDB dataset is very balanced, with 50% positive reviews and 50% negative reviews, the F-score of the models is very similar in all of the experiments using traditional classification algorithms such as NB, LR, and SVM, as shown in Table 2.

The IMDB movie review dataset and the Airlines Tweets dataset each have their own unique characteristics, and their performance varied when three fundamental architectures for deep learning: MLP, CNN, and LSTM-RNN were used. As the tests are run on a portable computer, the models are only trained for a total of five epochs so that the process can be completed in a shorter amount of time, as shown in Figure 5. Figure 5 plots accuracy vs. epoch and loss vs. epoch trends. In every epoch loss reduces and accuracy increases both for training and test dataset.

When applied to the IMDB dataset, the most accurate prediction that can be made using a typical machine learning technique (LR) is 77%, whereas the most accurate prediction that can be made using a deep learning algorithm (both CNN and LSTM-RNN) is 84%. When applied to tweets about airlines, the DL algorithm achieved the highest possible test accuracy of 91% using CNN architecture.

The outcomes of the experiment allow for a few different inferences to be made. Based on the results of the experiment, we can observe that the accuracy of both conventional algorithms and cutting-edge deep learning techniques improves as the size of the datasets and the number of features in each dataset grows (Table 2, Table 3). It is obvious that deep learning neural networks with CNN and bidirectional LSTM RNN achieve higher test accuracy (84%) than other traditional statistical machine learning algorithms and the simplest ANN architecture like MLP. In addition, applying deep learning to the dataset of airline tweets produced better results than applying it to the dataset of IMDB movie reviews. We concluded that airline tweets were more focused on the choice of words on sentiment expression, and as a result, they had greater accuracy, because airline tweets were shorter in length than those of movie reviews.

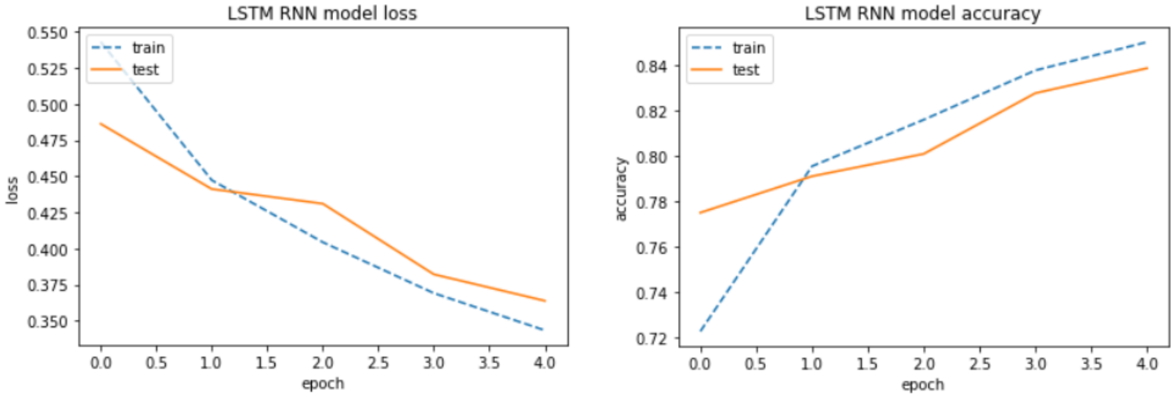

Figure 5. Accuracy vs. epoch (RHS) and loss vs. epoch (LHS) of LSTM-RNN (train vs. test dataset)

Table 2. Traditional machine learning algorithm comparison

IMDB dataset	NB F-Score	LR F-Score	SVM F-Score
2,000	0.71	0.71	0.68
20,000	0.76	0.76	0.63
30,000	0.75	0.75	0.67
50,000	0.76	0.77	0.72

Table 3. Deep learning architectures performance comparisons

IMDB-dataset size # of reviews	MLP test accuracy	CNN test accuracy	LSTM-RNN test accuracy
2,000	0.61	0.74	0.71
20,000	0.72	0.82	0.79
30,000	0.73	0.83	0.82
50,000	0.74	0.84	0.84
Airlines Tweets 15,000 reviews	0.88	0.91	0.83

5. CONCLUSION

The purpose of this paper was to understand the business impact of sentiment analysis in the big data era. In the future, there will be an increased demand for sentiment analysis tools in the fields of marketing, finance, political science, and health research since the amount of opinionated data available on social media will continue to expand. At the same time, developments in deep learning technologies, such as attention-based transformer architecture, are allowing for improved interpretation of the meanings of words and sentences. This paper discussed the challenges that come with translating natural languages like English into a format that a computer can understand automatically. The exploration of sentiment analysis for social media microtext presents extra obstacles, which are first covered, and then the current state of the art solution approach is given. The process of analyzing people's feelings remains difficult and calls for additional study efforts. In the future, the author intends to contribute to the development of tools for analyzing sentiments in low-resource Indian regional languages, such as those for which labeled datasets are difficult to come by. Although there are a great number of reviews and blogs written in regional languages these days, research on sentiment analysis in Indian regional languages is still not receiving a lot of attention. The readers of this work will be inspired to experiment with a contemporary deep learning strategy in search of improved accuracy in SA. A data scientist can expand their skill set by developing their own sentiment analysis tool using either standard machine learning algorithms or newer deep learning neural nets that can be tweaked using hyperparameters. This will allow the data scientist to analyze user feedback more accurately.

REFERENCES

[1] B. Liu, “Sentiment analysis: A multi-faceted problem,” *IEEE Intelligent Systems*, vol. 25, no. 3, pp. 76–80, 2010.

[2] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, “Sentiment strength detection in short informal text,” *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010, doi: 10.1002/asi.21416.

[3] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, “Sentiment analysis is a big suitcase,” *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, Nov. 2017, doi: 10.1109/MIS.2017.4531228.

[4] 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, doi: 10.1561/1500000011.

[5] K. P. K. and N. S., “Insights to problems, research trend and progress in techniques of sentiment analysis,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 5, p. 2818, Oct. 2017, doi: 10.11591/ijece.v7i5.pp2818-2822.

[6] R. D. and V. S., “Framework for opinion as a service on review data of customer using semantics based analytics,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 5, p. 5453, Oct. 2020, doi: 10.11591/ijece.v10i5.pp5453-5461, doi: 10.11591/ijece.v10i5.pp5453-5461.

[7] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, “A study on similarity and relatedness using distributional and wordnet-based approaches,” in *NAACL HLT 2009 - Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 2009, pp. 19–27, doi: 10.3115/1620754.1620758.

[8] P. D. Turney, “Thumbs up or thumbs down?,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02*, 2001, p. 417, doi: 10.3115/1073083.1073153.

[9] D. Gerz, I. Vulić, F. Hill, R. Reichart, and A. Korhonen, “SimVerb-3500: A large-scale evaluation set of verb similarity,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2173–2182, doi: 10.18653/v1/D16-1235.

[10] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, “Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques,” in *Third IEEE International Conference on Data Mining*, 2003, pp. 427–434, doi: 10.1109/ICDM.2003.1250949.

[11] R. Guha, “Exploring the field of text mining,” *International Journal of Computer Applications*, vol. 177, no. 4, pp. 11–17, Nov. 2017, doi: 10.5120/ijca2017915682.

[12] R. Guha, “Improving the performance of an artificial intelligence recommendation engine with deep learning neural nets,” in *2021 6th International Conference for Convergence in Technology (I2CT)*, Apr. 2021, pp. 1–7, doi: 10.1109/I2CT51068.2021.9417936.

[13] R. Guha, “Designing a chat-bot for college information using information retrieval and automatic text summarization techniques,” *Current Chinese Computer Science*, vol. 1, no. 1, pp. 42–51, May 2021, doi: 10.2174/2665997201999201022191540.

[14] B. Liu, *Sentiment Analysis*. Cambridge University Press, 2015.

[15] D. Jannach and M. Jugovac, “Measuring the business value of recommender systems,” *ACM Transactions on Management Information Systems*, vol. 10, no. 4, pp. 1–23, Dec. 2019, doi: 10.1145/3370082.

[16] C. A. Gomez-Uribe and N. Hunt, “The Netflix recommender system,” *ACM Transactions on Management Information Systems*, vol. 6, no. 4, pp. 1–19, Jan. 2016, doi: 10.1145/2843948.

[17] D. L. Pennell and Y. Liu, “Normalization of informal text,” *Computer Speech & Language*, vol. 28, no. 1, pp. 256–277, Jan. 2014, doi: 10.1016/j.csl.2013.07.001.

[18] G. Paltoglou and M. Thelwall, “A study of information retrieval weighting schemes for sentiment analysis,” in *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2010, pp. 1386–1395.

[19] A. Montoyo, P. Martínez-Barco, and A. Balahur, “Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments,” *Decision Support Systems*, vol. 53, no. 4, pp. 675–679, Nov. 2012, doi: 10.1016/j.dss.2012.05.022.

[20] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, “Comparing automated text classification methods,” *International Journal of Research in Marketing*, vol. 36, no. 1, pp. 20–38, Mar. 2019, doi: 10.1016/j.ijresmar.2018.09.009.

[21] C. Siebert, J. Hartmann, M. Heitmann, and C. Schamp, “Accuracy of automated sentiment analysis,” *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3489963.

[22] Y. Liu *et al.*, “RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv preprint arXiv:1907.11692*, Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692.

[23] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, vol. 1, pp. 142–150.

[24] R. Sukthanker, S. Poria, E. Cambria, and R. Thirunavukarasu, “Anaphora and coreference resolution: A review,” *Information Fusion*, vol. 59, pp. 139–162, Jul. 2020, doi: 10.1016/j.inffus.2020.01.010.

[25] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, Sep. 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391:AID-ASI1>3.0.CO;2-9.

[26] D. M. Blei, A. Y. Ng, and M. T. Jordan, “Latent dirichlet allocation,” *Advances in Neural Information Processing Systems*, vol. 3, no. Jan, pp. 993–1022, 2002.

[27] R. Guha, “Exploring Information retrieval by Latent Semantic Indexing and Latent Dirichlet Allocation Techniques,” *International Research Journal of Computer Science*, vol. 07, no. 05, pp. 54–65, May 2020, doi: 10.26562/irjcs.2020.v0705.001.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality arXiv : 1310 . 4546v1 [cs . CL] 16 Oct 2013,” *arXiv preprint arXiv:1310.4546*, vol. cs.CL, pp. 1–9, 2013, [Online]. Available: http://arxiv.org/abs/1310.4546v1%0Apapers2://publication/uuid/FB1742A3-202C-44CA-98F5-6EA51EC019D2.

[29] J. Pennington, R. Socher, and C. Manning, “Glove: global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[30] R. Socher *et al.*, “Recursive deep models for semantic compositionality over a sentiment treebank,” in *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2013, pp. 1631–1642.

[31] K. G. Kim, “Book Review: Deep Learning,” *Healthcare Informatics Research*, vol. 22, no. 4, p. 351, 2016, doi: 10.4258/hir.2016.22.4.351.

[32] Y. Goldberg, “A primer on neural network models for natural language processing,” *Journal of Artificial Intelligence Research*, vol. 57, pp. 345–420, Nov. 2016, doi: 10.1613/jair.4992.

[33] C. Dyer, A. Kuncoro, M. Ballesteros, and N. A. Smith, “Recurrent neural network grammars,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 199–209, doi: 10.18653/v1/N16-1024.

BIOGRAPHIES OF AUTHORS

Radha Guha received her Ph.D. degree in Computer Systems and Software from University of California Irvine in 2007. Her current research interest is in the area of application acceleration for the high-performance computation loads by the reconfigurable computing platform of Field Programmable Gate Arrays (FPGAs), artificial intelligence, machine learning and deep learning to solve emergent problems of the big data era like automatic text summarization, sentiment analysis, recommender system design, and face recognition. She is a faculty of Computer Science and Engineering at SRM Univ. AP., India. Her expertise is in artificial intelligence, machine learning, information retrieval, programming, soft computing and computer architecture. Before joining the academia, she was working in the industry as an Electrical Engineer and then as a Software Consultant. She can be contacted at email address: radhaguha9@gmail.com.

Tole Sutikno is currently employed as a lecturer in the Electrical Engineering Department at Universitas Ahmad Dahlan (UAD), which is in Yogyakarta, Indonesia. In 1999, 2004, and 2016, he graduated with a Bachelor of Engineering from Universitas Diponegoro, a Master of Engineering from Universitas Gadjah Mada, and a Doctor of Philosophy in Electrical Engineering from Universiti Teknologi Malaysia. All three degrees are in the field of electrical engineering. Since the year 2008, he has held the position of Associate Professor at the University of Agriculture and Development in Yogyakarta, Indonesia. He is currently the Head of the Embedded Systems and Power Electronics Research Group in addition to holding the position of Editor-in-Chief of TELKOMNIKA. His research interests include the areas of digital design, industrial applications, industrial electronics, industrial informatics, power electronics, motor drives, renewable energy, FPGA applications, embedded systems, artificial intelligence, digital libraries, and intelligent control. He can be contacted at email: tole@te.uad.ac.id.