Computer Assisted Methods in Engineering and Science
30(1): 7–26, 2023, doi: 10.24423/cames.471

This article belongs to the *Special Issue on Information Theory with Artificial Intelligence in 5G/6G Networks* edited by Dr. A.K. Budati, Dr. M.K. Hasan, Dr. R. Arthi, Dr. K. Thangavel

Handling Dimensionality of Ambiguity Using Ensemble Classification in Social Networks to Detect Multi-Label Sentiment Polarity

Sudarshan S. SONAWANE*, Satish R. KOLHE

School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon, India; e-mail: srkolhe2000@gmail.com

* *Corresponding Author e-mail: sudars2000@gmail.com*

With ever-increasing demand, social media platforms are rapidly developing to enable users to express and share their opinions on a variety of topics. Twitter is one such social media site. This platform enables a comprehensive view of the social media target setting, which may include products, social events, political scenarios, and administrative resolutions. The accessible tweets expressing the target audience's perspective are frequently impacted by ambiguity caused by natural language processing (NLP) limitations. By classifying tweets according to their sentiment polarity, we can determine whether they express a good or negative point of view, a neutral opinion, or an input tweet that is irrelevant to the sentiment polarity context. Categorizing tweets according to their sentiment can assist future activities within the target domain in constructively evaluating the sentiment polarity and enabling improved decision-making based on the observed sentiment polarity. In this study, tweets that were previously categorized with one of the sentiment polarities were used to conduct predictive analytics of the new tweet to determine its sentiment polarity. The ambiguity of the tweets corpus utilized in the training phase is a critical limitation of the sentiment categorization procedure. While several recent models proposed sentiment classification algorithms, they confined themselves to two labels: positive and negative opinion, oblivious to the plague of ambiguity in the training corpus. In this regard, a novel multi-label classification of sentiment polarity called handling dimensionality of ambiguity using ensemble classification (HAD-EC) method, which diffuses ambiguity and thus minimizes false alerts, is proposed. The experimental assessment validates the HAD-EC approach by comparing the suggested model's performance to other two existing models.

Keywords: sentiment analysis, ambiguity, fuzzy *c*-means, NLP, sentiment polarity, Twitter sentiment.

1. Introduction

Natural language processing (NLP) is a crucial area in data science because it allows computers to understand human language [1]. One is able to do this by using NLP evolved from computational linguistics. Computational linguistics is the name given to the cutting-edge study of linguistics that uses data science technologies [2]. To fully comprehend natural language, one must be well-versed in semantics, syntax, and a host of other disciplines.

We live in the era of high-speed networks such as 5G and big data, and as a result, we are constantly bombarded with data such as e-mails, SMS, online pages, and transactions. To find the valuable information hidden in this data, it must be evaluated. In this project, we focus on the brief text. Tweets, microblogs, and web searches are all examples of short texts. There are fewer statistics in such brief pieces of writing, making them more difficult to work with. Thus, being able to comprehend brief content can enhance our overall knowledge.

Sentiment analysis [3] is an important area in NLP since it analyzes and summarizes user opinions based on the collected data. Opinion mining, another term for sentiment analysis, is a newer study area.

Sentiment analysis and NLP are key components of the proposed system. If one has a problem with the system, one can enter it in the natural language and receive a timely answer in the same language it was entered. The sentiment analysis also calculates the strength of the user's complaint, which aids in allocating the correct priority among the users.

Recently, sentiment analysis has received a lot of interest in the NLP field. NLP can be used in a wide range of research areas since it can be used to mine people's thoughts and feelings. This topic has previously been tackled in three ways: opinion mining, phrase sentiment classifying, and word polarity prediction. Document and phrase sentiment analysis both rely largely on word-level systems for analysis.

Identifying the polarity of words is a common research topic [3–6], with most work focusing on isolating words or word senses from their context and assigning a prior polarity to them. However, the polarity of some words fluctuates significantly depending on the context, making it difficult to categorize them. In today's world, word disambiguation tasks take into account reducing noun sentiment classification. According to previous research, nouns can be classified as having a positive or negative sentiment polarity. By contrast, few other contemporary models [6–10] identify the sentiment polarity by associating nouns and adjectives. However, these approaches are infeasible to handle the curse of ambiguity in the training corpus, which leads to intolerable false alarm rate and minimal sensitivity and specificity. In addition, these contemporary contributions require highly accurate data sets for training, which is reasonably impossible to obtain. Moreover, word-based polarity assessment is not significant against a large volume of the training corpus, which often reflects the curse of ambiguity.

1.1. Problem statement

While social media networks such as Twitter allow users to express their views on a wide range of topics, they are particularly useful for reaching a broad audience in areas such as marketing and public policy. It is possible to determine the target audience's viewpoint using data science and data mining strategies such as sentiment analysis. Because of the uncertainty in target evaluations projected on social media platforms such as tweets on the Twitter network, sentiment polarity prediction or classification frequently fails with high false alert levels. Therefore, computer-aided sentiment prediction methods may lose their value as a contribution tool in this situation. Sentiment analysis's contributions from modern research show that ambiguity detection in input tweets is a critical problem of sentiment analysis.

1.2. Objective

A novel multi-label classification model is presented to predict the ambiguity in the tweets with diversified labels reflecting the respective sentiment polarity of the given training corpus.

2. Related work

Single-text interpretation faces many obstacles, including short text that is ambiguous and noisy, lack of statistical support for text mining approaches, and brief text message not always following written language grammar. Hua *et al.* [11] proposed an approach to short text interpretation incorporating semantic information from well-known database such as Probase. Their study involved online and offline parts. In the offline part, the authors built a language directory and collected data from a web archive and a central repository. Next, the semantic coherence between phrases to be used for online short text understanding was pre-calculated. In the online part, language segment, type categorization, and concept labeling were all used to generate the accurate interpretation of the short text.

Using a method developed by Lee *et al.* [12], the attribute may be extracted and then scored. Understanding the concept requires being able to identify key features. Concepts, entities, instances, characteristics, and relations make up the knowledge base. To extract qualities from various sources of data, the authors used the Probes knowledge base and combined concept-based and instance-based techniques.

NLP relies heavily on sentiment analysis. Text categorizing algorithms are used to analyze opinions and perspectives expressed in a text to conclude whether they are positive opinion, negative opinion, or neutral, which is known as the process of sentiment analysis (SA). In SA, each sentence is given a positive or negative polarity to determine the emotion expressed. Assigning polarity to the entire document (positive, negative, or neutral) is how SA is performed at the document level. At the aspect level, SA does a more in-depth analysis of the text, retrieving specific attributes and classifying viewpoints as positive or negative based on directly focusing on the opinions [13, 14].

The survey conducted in [13] examined the various sentiment analysis approaches, including linguistic, machine learning, and n-gram methodology. These techniques utilize a vocabulary that includes words representing both positive opinions and negative opinions, and matches it to the input text. There is no need for training data for this strategy, but it can be difficult to create a vocabulary that is unique and can be applied in many situations.

The disadvantage of machine learning technique is the difficulty in obtaining labeled data, which can be fairly costly [15–19] when training and developing new tasks.

The n-gram methods search for the text document's sequence for the n consecutive words. The value of n as 1 in n-gram denotes a unigram. Similarly the n value 2 and 3 denotes the n-grams as bigrams and trigrams, respectively. The n value 4 and above denotes n-grams by default [16]. The lexicon-based approach for sentiment classification [20] provides a way to use the Twitter API to analyze the sentiment of users. Laughing aloud is used in place of well-known acronym of LOL. Also, emojis are employed in order to determine the tweet's emotional content. Documents, aspects, and entities are all subjected to sentiment analysis. Machine learning for sentiment analysis in [15] develops a sentiment analysis technique handling anaphora dilemmas and using features to train the SVM classifier.

The other recent model-scale to estimate the aspects oriented sentiment polarity (SEAOSP) which is Sonawane and Kolhe [21] earlier contribution, successfully estimated the sentiment polarity by the correlation of the arguments and activities and their association with sentiment lexicons. Here, the arguments are considered as aspects and then these aspects' sentiment polarity are estimated, which is represented by the association of activities (adjectives) and sentiment lexicons. The experiments performed on standard datasets of tweets concluded that the SEAOSP outperformed the other contemporary models. However, the performance of SEAOSP was evaluated on standard datasets, which have high sensitivity and specificity and, collected tweets do not reflect ambiguity, which is a crucial constraint of the sentiment polarity assessment. Most of the contemporary models have not considered ambiguity as a critical constraint.

The other current model called ambiguity in sentiment classification (ABSC) aims to achieve minimal false alarming [22]. This ABSC approach uses dependency features to estimate sentiment polarity. However, extracting appropriate dependency features again has dimensionality constraint and needs a corpus with reasonable sensitivity and specificity to classify.

3. Methods and materials

This section describes the methods and materials used to handle the dimensionality of ambiguity by using ensemble classification from Twitter trends to detect multi-label sentiment polarity, as outlined in this article.

3.1. Model narrative

The initial phase performs data processing [3] to find a bag of words from each tweet of the given twitter corpus and further discovers the aspects to differentiate the words producing ambiguity. Afterward, the model discovers fuzzy clusters of the processed tweets using the fuzzy c-means algorithm, which conduct clustering by aspects as centroids discovered in the previous phase. Each resultant cluster includes a set of tweets that are the processed tweets with no ambiguity. Further, an ensemble classification to detect their sentiment polarity as positive, negative, neutral, or not-relevant is performed.

3.2. Preprocessing

Records with no label should be removed from the tweet corpus T. Preprocessing divides each record $\{r \,\exists\, r \in T\}$ into sentences, with each sentence containing the label "global polarity" from the original record. As a result of the corpus T, a list S of sentences is created. Next, each sentence $\{s \,\exists\, s \in S\}$ of the tweet $\{r \,\exists\, r \in T\}$, the preprocessing phase divides each tweet sentence $\{s \,\exists\, s \in S\}$ into a vector of tokens vt_s, each representing a word from the original sentence. Next, the tokens that indicate stop words from each vector are removed, and then the "ing" and "ed" forms of the remaining word tokens are removed. Each sentence $\{s \,\exists\, s \in S\}$ token vector vt_s is assigned to a set VT, and each element in the set represents a sentence $\{s \,\exists\, s \in S\}$ token vector vt_s.

3.3. The data structure

The tweets are presented as records of the given corpus $cT = \{t_1, t_2, ..., t_i, t_{i+1}, ..., t_{|cT|-1}, t_{|cT|}\}$. After data processing, the resultant corpus cP represents a set of records $cP = \left\{bw_1, bw_2, ..., bw_i, bw_{i+1}, ..., bw_{|cP|-1}, bw_{|cP|}\right\}$, which retains their sentiment polarity label given in the source corpus of tweets cT, such that each record $\{bw_i \,\exists\, bw_i \in cP \wedge 1 \leq i \leq |cP|\}$ represents the bag of words that are tokened from the record $\{t_i \,\exists\, t_i \in cT \wedge 1 \leq i \leq |cT| \wedge |cT| \equiv |cP|\}$.

3.4. Aspects discovery

The resultant corpus cP is used as input in next phase that discovers aspects as a set $cA = \left\{a_i, a_2, a_3, ..., a_{|cA|}\right\}$, which are the unique arguments discovered from each record $\{bw_i \,\exists\, bw_i \in cP \wedge 1 \leq i \leq |cP|\}$ of the bag of words of the set cP.

3.5. Ambiguity identification

The clustering technique adopted to remove the ambiguity of the given input tweets is fuzzy c-means [23]. The input tweets $\{t_i \,\exists\, t_i \in cT \wedge 1 \leq i \leq |cT|\}$ are processed and represented as the set cP of records $\{bw_i \,\exists\, bw_i \in cP \wedge 1 \leq i \leq |cP|\}$ representing the corpus of the bag of words of corresponding tweets $|cT|$ with their respective sentiment polarity.

The input data $\{bw_i \,\exists\, bw_i \in cP \wedge 1 \leq i \leq |cP|\}$ is divided into clusters by the fuzzy c-means method, with every cluster retaining a group of records with a substantial link.

Clustering is conducted by aspects as cluster centers (centroids), allowing each tweet into one or more clusters. The tweets are placed in a cluster by the distance between those tweets and the centroid of the respective cluster. A new centroid of each resulting cluster is found using Eqs. (1) and (2) and the process is repeated until the cluster centers do not change:

$$\underset{j=1}{\overset{|C_a|}{\forall}} \left\{ \mu_{ij} = 1 \Big/ \sum_{k=1}^{|cP|} \left(\frac{\left| c_a^j \cap bw_{ik} \right|}{\left| c_a^j \right|} \right)^{(2/fi-1)} \right\}, \tag{1}$$

$$\underset{j=1}{\overset{|C_a|}{\forall}} \left\{ c_a^j = \left\{ \left(\sum_{i=1}^{|cP|} (\mu_{ij})^{fi} * |bw_i| \right) \Big/ \left(\sum_{i=1}^{|cP|} (\mu_{ij})^{fi} \right) \right\} \right\}. \tag{2}$$

The notation $|cP|$ indicates the number of records representing a bag of words. The notation ca_j is the bag of words with less distance from the j-th cluster, and the index fuzziness is denoted by $fi \in [1, \infty]$. The set $C_a = \left\{c_a^1, c_a^2, ..., c_a^{|C_a|}\right\}$ represents the centroids. The notation μ_{ij} denotes the distance between the j-th cluster's centroid c_a^j and the bag of words $\{bw_i \,\exists\, bw_i \in cP \wedge 1 \leq i \leq |cP|\}$ representing the i-th record. The objective function J of the clustering technique follows Eq. (3):

$$J(U, V) = \sum_{i=1}^{|cP|} \sum_{j=1}^{|cA|} \left\{ (\mu_{ij})^{fi} \left\| \frac{\left| c_a^j \cap bw_i \right|}{\left| c_a^j \right|} \right\|^2 \exists\, i \leq |cP| \wedge j \leq |cA| \right\}. \tag{3}$$

$\frac{\left|c_a^j \cap bw_{ik}\right|}{\left|c_a^j\right|}$ // is the Euclidean distance of the j-th cluster's centroid c_a^j as well as the i-th record of the bag of words $\{bw_i \,\exists\, bw_i \in cP \wedge 1 \leq i \leq |cP|\}$.

The flow of the clustering process is as follows:
The set $cP = \left\{bw_1, bw_2, ..., bw_i, bw_{i+1}, ..., bw_{|cP|-1}, bw_{|cP|}\right\}$ represents a set of records such that each record $\{bw_i \,\exists\, bw_i \in cP \wedge 1 \leq i \leq |cP|\}$ represents the bag of words, whereas the notation $Ca = \left\{c_a^1, c_a^2, c_a^3, ..., c_a^{|Ca|}\right\}$ // indicates a set of centroids.

1) The cluster's centroid c_a^j of the j-th cluster is selected randomly.
2) The fuzzy membership μ_{ij} is computed by Eq. (4):

$$\mu_{ij} = 1/\sum_{k=1}^{|cP|}\left(\frac{\left|c_a^j \cap bw_{ik}\right|}{\left|c_a^j\right|}\right)^{(2/fi-1)}. \tag{4}$$

3) Equation (5) allows to select the centroid:

$$c_a^j = \left\{\left(\sum_{i=1}^{|cP|}(\mu_{ij})^{fi} * |bw_i|\right) \Big/ \left(\sum_{i=1}^{|cP|}(\mu_{ij})^{fi}\right)\right\}. \tag{5}$$

4) Fuzzy membership is computed using Eq. (4), and centroid is found using Eq. (5) until $\beta \geq \|U(k) - U(k+1)\|$, where k and β are the index of the current iteration and termination criteria, respectively. The expression $U = |Ca| * (|cP| * (\mu_{ij}))$ indicates a two-dimensional vector of fuzzy membership.

Let the set $fC = \left\{fc_1, fc_2, ..., fc_{|fC|}\right\}$ denote the set of resultant clusters.

3.6. The classification model

3.6.1. The fitness function of the training phase. This section explains the process of estimating confidence coefficients of aspects and lexicons toward positive and negative labels of each cluster $\{fc_i \,\exists\, fc_i \in fC \wedge 1 \leq i \leq |fC|\}$ discovered using the fuzzy c-means method, which has been explained in the earlier section. Model of the estimating confidence coefficients of aspects and lexicons toward positive and negative labels is as follows:

$\overset{|fC|}{\underset{j=1}{\forall}} \{fc_j \,\exists\, fc_j \in fC \wedge j \leq |fC|\}$ Begin // for each cluster

$\{fc_i \,\exists\, fc_i \in fC \wedge 1 \leq i \leq |fC|\}$.

Aspect's positive normalized confidence:

$$\underset{i=1}{\overset{|fc_j|}{\forall}} \left\{bw_i^+ \,\exists\, bw_i^+ \in fc_j \wedge i \leq |fc_j|\right\},$$

$$apnc(bw_i^+) = 1 - \frac{1}{\left|\left(bw_i^+ \cap cA\right)\right|}.$$

Aspect's negative normalized confidence:

$$\underset{i=1}{\overset{|fc_j|}{\forall}} \left\{bw_i^- \,\exists\, bw_i^- \in fc_j \wedge i \leq |fc_j|\right\},$$

$$annc(bw_i^-) = 1 - \frac{1}{\left|\left(bw_i^- \cap cA\right)\right|}.$$

Positive lexicon's normalized confidence:

$$\underset{i=1}{\overset{|fc_j|}{\forall}} \left\{bw_i^+ \,\exists\, bw_i^+ \in fc_j \wedge i \leq |fc_j|\right\},$$

$$plnc(bw_i^+) = 1 - \frac{1}{\left|\left(bw_i^+ \cap Lx_+\right)\right|}.$$

Negative lexicon's normalized confidence:

$$\underset{i=1}{\overset{|fc_j|}{\forall}} \left\{bw_i^- \,\exists\, bw_i^- \in fc_j \wedge i \leq |fc_j|\right\},$$

$$nlnc(bw_i^-) = 1 - \frac{1}{\left|\left(bw_i^- \cap Lx_-\right)\right|}.$$

Aspect's positive confidence coefficient:

$$\langle apnc\rangle = \frac{\sum_{i=1}^{|fc_j|} apnc\left(bw_i^+\right)}{\left|fc_j^+\right|},$$

$$apcc_j = \langle apnc\rangle - \left(\frac{\sum_{i=1}^{|fc_j|} \left\{\left\|\langle apnc\rangle - apnc\left(bw_i^+\right)\right\| \,\exists\, bw^+ \in fc_j\right\}}{\left|fc_j^+\right|}\right).$$

Aspect's negative confidence coefficient:

$$\langle annc\rangle = \frac{\sum_{i=1}^{|fc_j|} \left\{annc\left(bw_i^-\right) \,\exists\, bw_i^- \in fc_j\right\}}{\left|fc_j^-\right|},$$

$$ancc_j = \langle annc \rangle - \left(\frac{\sum_{i=1}^{|fc_j|} \left\{ \left\| \langle annc \rangle - annc\left(bw_i^-\right) \right\| \exists\, bw_i^- \in fc_j \right\}}{\left| fc_j^- \right|} \right).$$

Lexicon's positive confidence coefficient:

$$\langle plnc \rangle = \frac{\sum_{i=1}^{|fc_j|} \left\{ plnc(bw_i^+) \,\exists\, bw_i^+ \in fc_j \right\}}{|fc_j|},$$

$$plcc_j = \langle plnc \rangle - \left(\frac{\sum_{i=1}^{|fc_j|} \left\{ \left\| \langle plnc \rangle - plnc\left(bw_i^+\right) \right\| \exists\, bw_i^+ \in fc_j \right\}}{|fc_j|} \right).$$

Lexicon's negative confidence coefficient:

$$\langle nlnc \rangle = \frac{\sum_{i=1}^{|fc_j|} \left\{ nlnc(bw_i^-) \,\exists\, bw_i^- \in fc_j \right\}}{|fc_j|},$$

$$nlcc_j = \langle nlnc \rangle - \left(\frac{\sum_{i=1}^{|fc_j|} \left\{ \left\| \langle nlnc \rangle - nlnc\left(bw_i^-\right) \right\| \exists\, bw_i^- \in fc_j \right\}}{|fc_j|} \right).$$

3.6.2. Prediction phase. For each given unlabeled tweet t, the label prediction process shall perform as follows.

Apply tweet processing on a given tweet as described in Subsec. 3.1, which results in the bag of words bw. Further, extract all arguments existing in the bag of words bw as aspects aT. Afterward, find all lexicons existing in a given tweet as tL. Next, discover aspect and lexicon confidence of the given tweet toward all clusters $fC = \left\{ fc_1, fc_2, ..., fc_{|fC|} \right\}$ as follows:

$$\forall_{j=1}^{|fC|} \left\{ fc_j \,\exists\, fc_j \in fC \wedge j \leq |fC| \right\} \text{ Begin // for each cluster}$$

$$\left\{ fc_i \,\exists\, fc_i \in fC \wedge 1 \leq i \leq |fC| \right\},$$

$\overset{|fc_j|}{\underset{i=1}{\forall}} \left\{ ua_j^+(t) \leftarrow \left(ua_j^+(t) \cap \left(tA \cap bw_i^+ \right) \right) \exists\, bw_i^+ \in fc_j \right\}$ // finding all unique positive aspects ua_j^+ of the tweet t in the j-th cluster fc_j,

$apnc_j(t) = 1 - \frac{1}{\left| ua_j^+(t) \right|}$ // finding the aspect's positive normalized confidence $apnc_j(t)$ of the given tweet t in the j-th cluster fc_j,

$\overset{|fc_j|}{\underset{i=1}{\forall}} \left\{ ua_j^-(t) \leftarrow \left(ua_j^-(t) \cap \left(tA \cap bw_i^- \right) \right) \exists\, bw_i^- \in fc_j \right\}$ // finding all unique negative aspects ua_j^- of the tweet t in the j-th cluster fc_j,

$annc_j(t) = 1 - \frac{1}{\left| ua_j^-(t) \right|}$ // finding the aspect's negative normalized confidence $annc_j(t)$ of the given tweet t in the j-th cluster fc_j,

$\overset{|fc_j|}{\underset{i=1}{\forall}} \left\{ ul_j^+(t) \leftarrow \left(ul_j^+(t) \cap \left(tL \cap bw_i^+ \right) \right) \exists\, bw_i^+ \in fc_j \right\}$ // finding all unique positive lexicons ul_j^+ of the tweet t in the j-th cluster fc_j,

$plnc_j(t) = 1 - \frac{1}{\left| ul_j^+(t) \right|}$ // finding the positive lexicon's normalized confidence $plnc_j(t)$ of the given tweet t in the j-th cluster fc_j,

$\overset{|fc_j|}{\underset{i=1}{\forall}} \left\{ ul_j^-(t) \leftarrow \left(ul_j^-(t) \cap \left(tL \cap bw_i^- \right) \right) \exists\, bw_i^- \in fc_j \right\}$ // finding all unique negative lexicons ul_j^- of the tweet t in the j-th cluster fc_j,

$nlnc_j(t) = 1 - \frac{1}{\left| ul_j^-(t) \right|}$ // finding the negative lexicon's normalized confidence $nlnc_j(t)$ of the given tweet $t$in the j-th cluster fc_j.

End

//Estimating the normalized weights of the aspects and lexicons of both positive and negative labels//

$\overset{|fC|}{\underset{j=1}{\forall}} \{fc_j \exists fc_j \in fC \wedge j \leq |fC|\}$ Begin // for each cluster

$\{fc_i \exists fc_i \in fC \wedge 1 \leq i \leq |fC|\}$,

$pa_j(t) = \left\{ \begin{array}{l} 1 \exists apnc_j(t) > apcc_j \\ 0 \end{array} \right\}$ // finding the positive aspects state $pa_j(t)$, which is 0 or 1 of the j-th cluster fc_j, such that the positive aspect's state $pa_j(t)$ of the given tweet t is 1 if the aspect's positive normalized confidence $apnc(t)$ of the given tweet t is greater than the aspect's positive confidence coefficient $apcc_j$ of the j-th cluster fc_j; otherwise the positive aspects state $pa_j(t)$ of the given tweet t is 0,

$na_j(t) = \left\{ \begin{array}{l} 1 \exists annc_j(t) > ancc_j \\ 0 \end{array} \right\}$ // finding the negative aspects state $na_j(t)$, which is 0 or 1 of the j-th cluster fc_j, such that the negative aspect's state $na_j(t)$ of the given tweet t is 1 if the aspect's negative normalized confidence $annc(t)$ of the given tweet t is greater than the aspect's negative confidence coefficient $ancc_j$ of the j-th cluster fc_j; otherwise the negative aspects state $na_j(t)$ of the given tweet$t$ is 0,

$pl_j(t) = \left\{ \begin{array}{l} 1 \exists plnc_j(t) > plcc_j \\ 0 \end{array} \right\}$ // finding the positive lexicons state $pl_j(t)$, which is 0 or 1 of the j-th cluster fc_j, such that the positive lexicon's state $pl_j(t)$ of the given tweet t is 1 if the positive lexicon's normalized confidence $plnc(t)$ of the given tweet t is greater than the positive lexicon's confidence coefficient $plcc_j$ of the j-th cluster fc_j; otherwise the positive lexicons state $pl_j(t)$ of the given tweet t is 0,

$nl_j(t) = \left\{ \begin{array}{l} 1 \exists nlnc_j(t) > nlcc_j \\ 0 \end{array} \right\}$ // finding the negative lexicons state $nl_j(t)$, which is 0 or 1 of the j-th cluster fc_j, such that the negative lexicon's state $na_j(t)$ of the given tweet t is 1 if the negative lexicon's normalized confidence $nlnc(t)$ of the given tweet t is greater than the negative lexicon's confidence coefficient $nlcc_j$ of the j-th cluster fc_j; otherwise the negative lexicon state $nl_j(t)$ of the given tweet t is 0,

End

//Counting the number of clusters with a positive and negative state of aspects and labels of the given tweet $t$in the normalized format//

$c^a_{+ve}(t) = 1 - \frac{1}{\sum_{j=1}^{|fC|}\{pa_j(t)\}}$ // the aggregate of clusters c^a_{+ve} with a positive state of the aspects as one in normal form,

$c^a_{-ve}(t) = 1 - \frac{1}{\sum_{j=1}^{|fC|}\{na_j(t)\}}$ // the aggregate of clusters c^a_{-ve} with a negative state of the aspects as one in normal form,

$c^l_{+ve}(t) = 1 - \frac{1}{\sum_{j=1}^{|fC|}\{pl_j(t)\}}$ // the aggregate of clusters c^l_{+ve} with a positive state of the lexicons as one in normal form,

$c^l_{-ve}(t) = 1 - \frac{1}{\sum_{j=1}^{|fC|}\{nl_j(t)\}}$ // the aggregate of clusters c^l_{-ve} with a negative state of the lexicons as one in normal form,

$w_{+ve}(t) = 1 - \left(c^a_{+ve}(t) \times c^l_{+ve}(t)\right)$ // finding the positive weight $w_{+ve}(t)$ of the given tweet t, which is the normalized product of the aggregate of clusters c^a_{+ve} with a positive state of the aspects and the aggregate of clusters c^l_{+ve} with a positive state of the lexicons of the given tweet t,

$w_{-ve}(t) = 1 - \left(c^a_{-ve}(t) \times c^l_{-ve}(t)\right)$ // finding the negative weight $w_{-ve}(t)$ of the given tweet t, which is the normalized product of the aggregate of clusters c^a_{-ve} with a negative state of the aspects and the aggregate of clusters c^l_{-ve} with a negative state of the lexicons of the given tweet t.

//Label Assessment//

$$\langle p\tau \rangle = \left\{ \begin{array}{l} \left(1 - \left(\sum_{j=1}^{|fC|}\{apcc_j\}\right)^{-1}\right) + \left(1 - \left(\sum_{j=1}^{|fC|}\{ancc_j\}\right)^{-1}\right) + \\ \left(1 - \left(\sum_{j=1}^{|fC|}\{plcc_j\}\right)\right) + \left(1 - \left(\sum_{j=1}^{|fC|}\{nlcc_j\}\right)^{-1}\right) \end{array} \right\}$$

$\times \left(2\left(\frac{\log^4}{\log^2}\right)\right)^{-1}$ // finding the mean of the aggregate of all confidence coefficients of both aspects and labels of all the clusters,

$$\sigma = \left\{ \left\{ \begin{array}{l} \left\| \langle p\tau \rangle - \left(1 - \frac{1}{\sum_{j=1}^{|fC|} \{apcc_j\}} \right) \right\| + \left\| \langle p\tau \rangle - \left(1 - \frac{1}{\sum_{j=1}^{|fC|} \{ancc_j\}} \right) \right\| + \\ \left\| \langle p\tau \rangle - \left(1 - \frac{1}{\sum_{j=1}^{|fC|} \{plcc_j\}} \right) \right\| + \left\| \langle p\tau \rangle - \left(1 - \frac{1}{\sum_{j=1}^{|fC|} \{nlcc_j\}} \right) \right\| \end{array} \right\} \times \left(2 \left(\frac{\log^4}{\log^2} \right) \right)^{-1} \right\} //$$

finding the deviation of all confidence coefficients of both aspects and labels of all the clusters,

$p\tau = \langle p\tau \rangle + \sigma$ // finding the label probability threshold $p\tau$, which is the sum of the mean value $\langle p\tau \rangle$ and the deviation,

$$\left\{ \begin{array}{l} lbl(t) = positive \, \exists \\ \left(c^a_{+ve}(t) - c^a_{-ve}(t) \right) > p\tau \wedge \\ \left(c^l_{+ve}(t) - c^l_{-ve}(t) \right) > p\tau \wedge \\ \left(w_{+ve}(t) - w_{-}(t) \right) > p\tau \wedge \\ \left(c^l_{+ve}(t) - c^l_{-ve}(t) \right) > p\tau \end{array} \right\} //$$

verifying whether the label scope *lbl(t)* is positive or not,

$$\left\{ \begin{array}{l} lbl(t) = negative \, \exists \\ \left(c^a_{-ve}(t) - c^a_{+ve}(t) \right) > p\tau \wedge \\ \left(c^l_{-ve}(t) - c^l_{+ve}(t) \right) > p\tau \wedge \\ \left(w_{-ve}(t) - w_{+}(t) \right) > p\tau \end{array} \right\} //$$

verifying whether the label scope *lbl(t)* is negative or not,

$$\left\{ \begin{array}{l} lbl(t) = neutral \, \exists \\ \left(c^l_{+ve}(t) - c^l_{-ve}(t) \right) < p\tau \wedge \\ \left(c^l_{-ve}(t) - c^l_{+ve}(t) \right) < p\tau \end{array} \right\} //$$

verifying whether the label scope *lbl(t)* is neutral or not,

$$\left\{ \begin{array}{l} lbl(t) = not - relevant \, \exists \\ \left(c^a_{+ve}(t) - c^a_{-ve}(t) \right) < p\tau \wedge \\ \left(c^a_{-ve}(t) - c^a_{+ve}(t) \right) < p\tau \end{array} \right\} //$$

verifying if the label scope *lbl(t)* is not-relevant or not.

4. Experimental study

The Apple Twitter sentiment data corpus [24] is used, which is the collection of tweets reflecting curse of ambiguity and consisting of 3887 tweets reflecting sentiment polarity labels: neutral, not-relevant, positive, and negative. The statistics of the data are presented in Table 1.

Table 1. Data statistics used for training and testing.

Label	Total	Train	Test
Negative	1458	1094	364
Neutral	1009	757	252
Positive	822	617	205
Not-relevant	598	449	149
Total	3887	2917	970

The performance of the proposed HAD-EC model and two other existing models of SEAOSP [21] and ABSC [22] is examined in the experimental investigation using the metrics of precision, sensitivity, specificity, F-measure, and accuracy over multi-label four-fold cross-validation. A further comparison is conducted between the proposed model and the current models.

4.1. Four-fold cross-validation

Figure 1 demonstrates precision represented by y-coordinates and four folds represented by x-coordinates, each with a different label such as positive, negative, neutral, or not-relevant. These labels are compared between the proposed HAD-EC model and the ABSC and SEAOSP models.

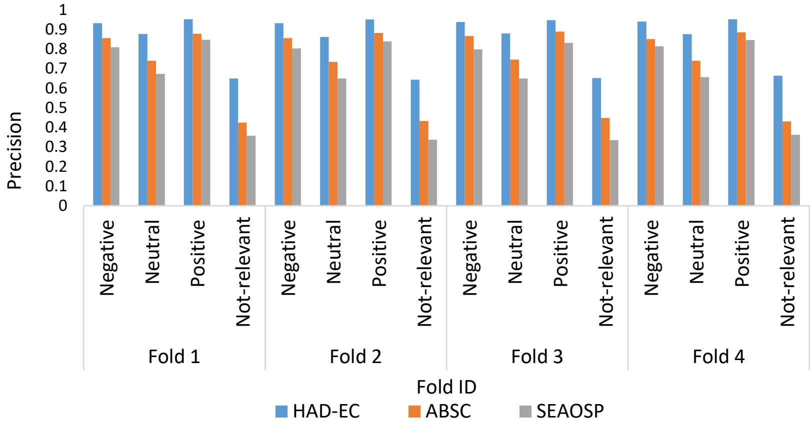

Fig. 1. Graphical depiction of precision across the four- fold comparison of the proposed HAD-EC model HAD-EC with current ABSC and SEAOSP models.

Sensitivity is also known as recall (ratio of true positives to total positives). This recall criterion measures the performance of the proposed HAD-EC and ABSC and SEAOSP models in the four-fold comparison, as shown in Fig. 2. Three models are compared in the graphs with four labels: positive, negative, neutral, and not-relevant.

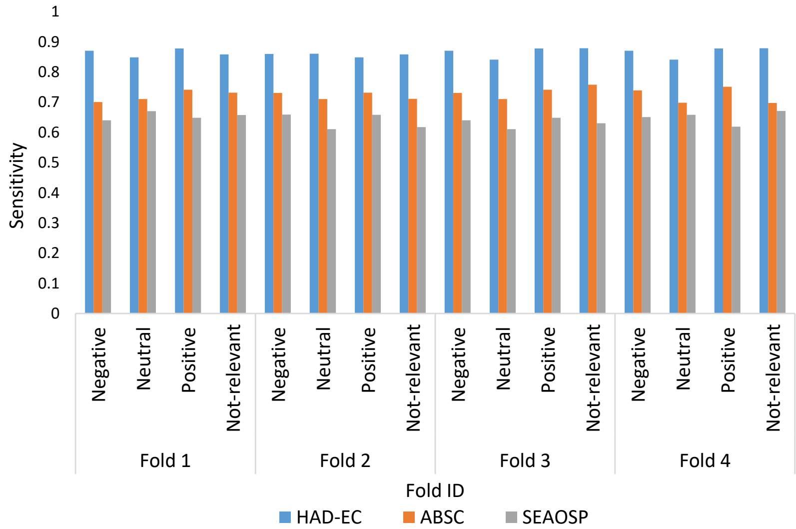

Fig. 2. Graphical depiction of sensitivity across the four-fold comparison of the proposed HAD-EC system with ABSC and SEAOSP models.

Specificity is another key element identified for performance assessment throughout the fourfold comparison of the three models evaluated in the investigation. Technically known as TNR true-negative rate (TNR), it is the ratio of TNs, where the total of FPs and TNs is evaluated as essential to the process. In terms of assessing the performance of the HAD-EC model and ABSC and SEAOSP models, the model performance as represented in Fig. 3 relates to how various models fared in the experimental investigation. Based on the inputs evaluated, it is clear that the solution provided in this paper, known as HAD-EC, performed well in contrast to the other two models.

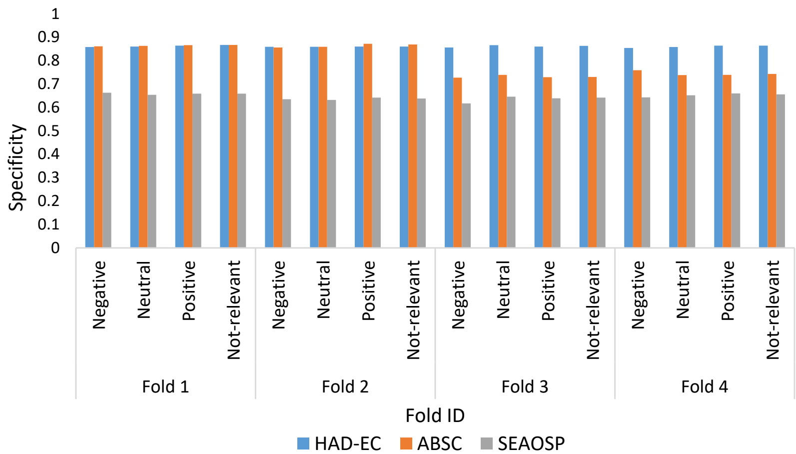

Fig. 3. Model performance for TNR specificity across diversity threshold values.

Figure 4 depicts a graph with an F-score represented by y-coordinates and four folds by x-coordinates, each with a different label such as positive, negative, neutral, or not-relevant. These labels are compared between the proposed HAD-EC model and ABSC and SEAOSP models.

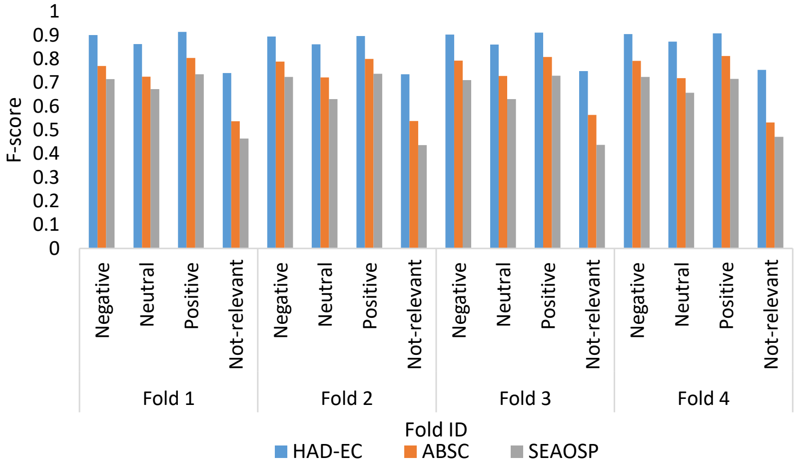

Fig. 4. Graphical depiction of the F-score four-fold comparison between the proposed HAD-EC model and ABSC and SEAOSP models.

The prediction accuracy measures the performance of the proposed HAD-EC model and ABSC and SEAOSP models using the four-fold comparison, as shown in Fig. 5. Three models are compared in the graph with four labels: positive, negative, neutral, and not-relevant.

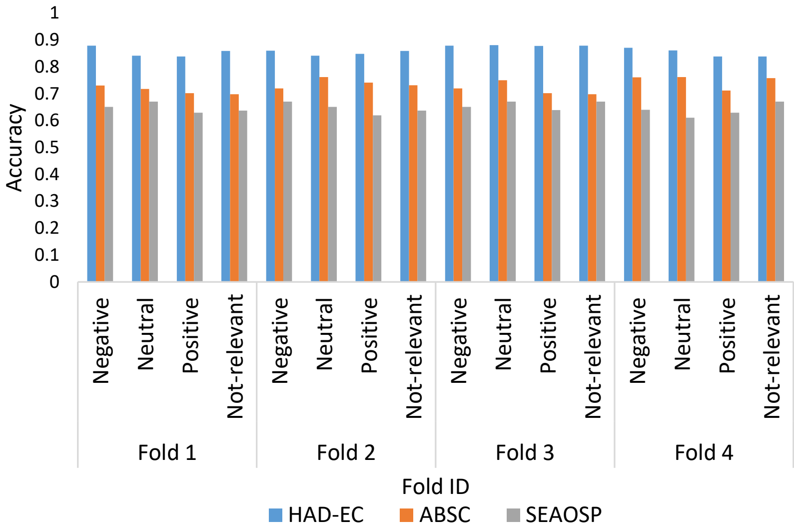

Fig. 5. Graphical depiction of accuracy in the four-fold comparison between the proposed HAD-EC model and ABSC and SEAOSP models.

Figure 6 depicts a graph plotting micro values of various metrics such as precision, sensitivity, specificity, F-score, and decision accuracy across four-fold comparison of the proposed HAD-EC model and ABSC and SEAOSP models. According to the above data depicted in this graph, the performance of the suggested model in all measures outperforms that of the other two models.

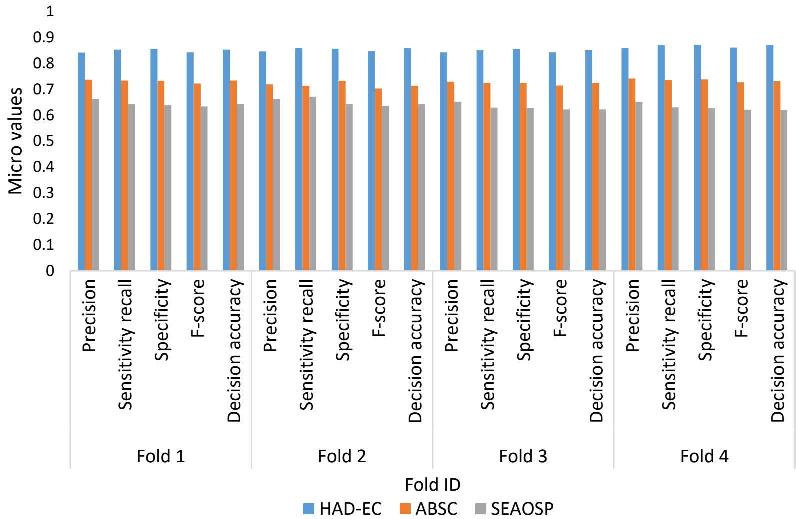

Fig. 6. A graphical depiction of the four-fold comparison of all metrics for the proposed HAD-EC system HAD-EC with ABSC and SEAOSP models.

5. Conclusion

To overcome the curse of ambiguity in a particular twitter corpus, it is frequently critical to determine the sentiment orientation of intended target evaluations or opinions. The purpose of this article was to describe how to overcome the sentiment analysis's curse of ambiguity. The objective of this contribution was to conduct ensemble learning and sentiment polarity prediction on a given corpus by using the fuzzy c-means method to split it into multiple fuzzy clusters. The fuzzy c-means clustering and ensemble classification techniques were used to perform sentiment polarity identification based on aspects. In contrast to other current techniques, the suggested fuzzy clustering splits the provided corpus into numerous groups, with each cluster containing one or more records. When individuals employ ambiguous phrases, they frequently associate them with a variety of possible meanings. As a result, clusters are defined by centroids. In this study, an ensemble supervised learning model is presented to determine the sentiment polarity of multiple labels classified as positive, negative, neutral, and not-relevant. The study examined a corpus of Apple tweets.

The cross-validation metrics produced from the ABSC and SEAOSP models were compared to those predicted by the HAD-EC model. According to the findings of the performance metrics, the proposed model (precision 0.83, accuracy 0.85, sensitivity 0.86, specificity 0.88, F-score 0.81) outperforms the existing ABSC model (precision 0.73, accuracy 0.72, sensitivity 0.74, specificity 0.73, F-score 0.76), and the SEAOSP model (precision 0.68, accuracy 0.65, sensitivity 0.64, specificity 0.66, F-score 0.62) in predicting the sentiment polarity in ambiguous social media data. It is feasible that future research may use this technique to determine the emotion of tweets using multi-dimensional components such as emoticons, emojis, and other sentiment representations.

Acknowledgements

This research work was carried at the School of Computer Sciences (SOCS), K.B.C.N.M.U., Jalgaon, under the SAP DRS-II level and was supported by UGC, New Delhi.

References

1. S. Wiegreffe, A. Marasović, Teach me to explain: A review of datasets for explainable natural language processing, [in:] *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 23 pages, 2021, https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/698d51a19d8a121ce581499d7b701668-Paper-round1.pdf.
2. Y. Belinkov, S. Gehrmann, E. Pavlick, Interpretability and analysis in neural NLP, [in:] *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 1–5, 2020, doi: 10.18653/v1/2020.acl-tutorials.1.
3. S.S. Sonawane, S.R. Kolhe, Feature optimization in sentiment analysis by term co-occurrence fitness evolution (TCFE), *International Journal of Information Technology and Web Engineering* (IJITWE), **14**(3): 16–36, 2019, doi: 10.4018/IJITWE.2019070102.
4. S.S. Sonawane, S.R. Kolhe, Term co-occurrence based feature selection for sentiment classification, [in:] P. Bhattacharyya, H. Sastry, V. Marriboyina, R. Sharma [Eds.], *Smart and Innovative Trends in Next Generation Computing Technologies, NGCT-2017*, Communications in Computer and Information Science, Springer, Vol. 827, pp. 405–417, 2018, doi: 10.1007/978-981-10-8657-1_31.
5. S.M. Rezaeinia, R. Rouhollah, A. Ghodsi, H. Veisi, Sentiment analysis based on improved pre-trained word embeddings, *Expert Systems with Applications*, **117**: 139–147, 2019, doi: 10.1016/j.eswa.2018.08.044.
6. F. Ali *et al.*, Transportation sentiment analysis using word embedding and ontology-based topic modeling, *Knowledge-Based Systems*, **174**: 27–42, 2019, doi: 10.1016/j.knosys.2019.02.033.
7. A. Onan, Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks, *Concurrency and Computation: Practice and Experience*, **33**(23): e5909, 2020, doi: 10.1002/cpe.5909.

8. C.S.G. Khoo, S.B. Johnkhan, Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons, *Journal of Information Science*, **44**(4): 491–511, 2018, doi: 10.1177/0165551517703514.

9. S. Bandari, V.V. Bulusu, Survey on ontology-based sentiment analysis of customer reviews for products and services, [in:] K. Raju, R. Senkerik, S. Lanka, V. Rajagopal [Eds.], *Data Engineering and Communication Technology*, Advances in Intelligent Systems and Computing, Springer, Vol. 1079, pp. 91–101, 2020, doi: 10.1007/978-981-15-1097-7_8.

10. C. Song, X.-K. Wang, P.-F. Cheng, J.-Q. Wang, L. Li, SACPC: A framework based on probabilistic linguistic terms for short text sentiment analysis, *Knowledge-Based Systems*, **194**, 2020, doi: 10.1016/j.knosys.2020.105572.

11. W. Hua, Z. Wang, H. Wang, K. Zheng, X. Zhou, Understand short texts by harvesting and analyzing semantic knowledge, *IEEE Transaction on Knowledge and Data Engineering*, **29**(3): 499–512, 2017, doi: 10.1109/TKDE.2016.2571687.

12. T. Lee, Z. Wang, H. Wang, S.-W. Hwang, Attribute extraction and scoring: A probabilistic approach, [in:] *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 194–205, Brisbane, QLD, Australia, April 8–12, 2013, doi: 10.1109/ICDE.2013.6544825.

13. A. D'Andrea, F. Ferri, P. Grifoni, T. Guzzo, Approaches, tools and applications for sentiment analysis implementation, *International Journal of Computer Applications*, **125**(3): 26–33, 2015, doi: 10.5120/ijca2015905866.

14. A. Hassan, A. Abbasi, D. Zeng, Twitter sentiment analysis: A bootstrap ensemble framework, [in:] *2013 IEEE International Conference on Social Computing*, pp. 357–364, Alexandria, VA, USA, September 8–14, 2013, doi: 10.1109/SocialCom.2013.56.

15. M.H. Krishna, K. Rahamathulla, A. Akbar, A feature based approach for sentiment analysis using SVM and coreference resolution, [in:] *IEEE International Conference on Inventive Communication and Computational Technologies(ICICCT)*, Coimbatore, India, March 10–11, 2017, doi: 10.1109/ICICCT.2017.7975227.

16. A. Tripathy, A. Agrawal, S.K. Rath, Classification of sentiment reviews using n-gram machine learning approach, *Expert Systems with Applications*, **57**: 117–126, 2016, doi: 10.1016/j.eswa.2016.03.028.

17. D. Zhang, H. Xu, Z. Su, Y. Xu, Chinese comments sentiment classification based on word2vec and $\mathrm{SVM}^{\mathrm{perf}}$, *Expert Systems with Applications*, **42**(4): 1857–1863, 2015, doi: 10.1016/j.eswa.2014.09.011.

18. B. Luo, J. Zeng, J. Duan, Emotion space model for classifying opinions in stock message board, *Expert Systems with Applications*, **44**: 138–146, 2016, doi: 10.1016/j.eswa.2015.08.023.

19. H. Kang, S.J. Yoo, D. Han, Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews, *Expert System Applications*, **39**(5): 6000–6010, 2012, doi: 10.1016/j.eswa.2011.11.107.

20. P. Ray, A. Chakrabarti, Twitter sentiment analysis for product review using lexicon method, [in:] *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, pp. 211–216, Pune, India, February 24–26, 2017, doi: 10.1109/ICDMAI.2017.8073512.

21. S.S. Sonawane, S.R. Kolhe, A new method for defining scale to estimate the aspects oriented sentiment polarity of the tweets, [in]: K.C. Santosh, B. Gawali [Eds.], *Recent Trends in Image Processing and Pattern Recognition (RTIP2R) 2020*, Communications in Computer and Information Science (CCIS), Springer, Vol. 1380, pp. 318–333, 2021, doi: 10.1007/978-981-16-0507-9_28.

22. S. Deng, A.P. Sinha, H. Zhao, Resolving ambiguity in sentiment classification: The role of dependency features, *ACM Transactions on Management Information Systems* (TMIS), **8**(2–3): 13 pages, 2017, doi: 10.1145/3046684.

23. M. Trupthi, S. Pabboju, G. Narsimha, Possibilistic fuzzy C-means topic modelling for Twitter sentiment analysis, *International Journal of Intelligent Engineering and Systems*, **11**(3): 100–108, 2018, doi: 10.22266/ijies2018.0630.11.

24. Apple Twitter Sentiment, online: https://data.world/crowdflower/apple-twitter-sentiment, accessed on 21.10.2021.

Received January 15, 2022; revised version March 3, 2022; accepted March 25, 2022.