Prediction of User-Brand Associations Based on Sentiment Analysis

Mariella Bonomo[1], Simona E. Rombo[1] and Filippo Rotolo[1]

[1]Department of Mathematics and Computer Science, University of Palermo

Abstract

Finding the right users to be chosen as targets for advertising campaigns is not a trivial task, and it may allow important commercial advantages. A novel approach is presented here for the recommendation of new possible consumers to brands interested in distributing advertising campaigns, ranked according to the "compatibility" between users and brands. A database containing both descriptions associated with different brands, and textual information about users' opinions on different topics, is required in input. Then, sentiment analysis techniques are applied to measure to what extent the users match with the brands, based on the texts associated with their opinions. The approach has been tested on both synthetic and real datasets, and with two different formulations, showing promising results in all the considered experiments.

Keywords

social advertising, social networks, sentiment analysis, user-brand associations

1. Introduction

An important issue in the context of digital advertising is how to optimize the effects of marketing communication, trying to involve in advertising campaigns those potential consumers who are the most interested ones, and avoiding the distribution of advertisements to uninterested users. Automatic systems able to suggest a set of target users for advertising campaigns, possibly ranked according to their potential approval rating, provide three main benefits: (i) minimization of costs for the dissemination of the advertising campaign through digital media, which is often very expensive; (ii) improvement of the user experience, since only the possibly interested customers are contacted with advertisements which could be useful for them; (iii) avoid the spread of useful information through the social and other digital channels.

Here we propose a novel approach for the recommendation of a list of possible consumers to be suggested as target for a specific advertisement campaign, ranked according to their "compatibility" with the brand promoting the campaign. In particular, we assume that a database containing a number of descriptions associated with different brands is available, and a number of potential customers together with their opinions on different topics as well. Suitable "tag" may be generated to summarize the brand descriptions, and the proposed approach is based on the adoption of sentiment analysis techniques [1, 2, 3, 4] to understand if users are compatible or not with the brands, based on the texts associated to their opinions. The proposed approach is the core of a more general framework, implementing a big data analytics platform.

The presented methodology has been validated first on synthetic data, according to the Receiver Operating Characteristic (ROC) analysis [5] which has shown an Area Under the Curve (AUC) equal to 0.975 in ideal case, and to 0.84 when a small percentage of noise is injected in the dataset. Then, a database of brands and possible customer features has been built, using web pages of some real brands and data retrieved from the Twitter social network [6, 7]. In particular, two different sets of experiments have been performed on this latter database. First, the accuracy of the method in correctly associating users that are "followers" of some brands with those brands, based only on their tweets and on the brand tags, have been evaluated, showing that the proposed approach has been able to correctly associate the 93.7% of the considered followers. Then, the ability of the method in classifying potential new customers has been tested by applying the K-Nearest Neighbors approach, which has returned the best result of 85.71% correctly classified users for $K = 6$.

2. Related Work

The authors of [8] use Differential Language Analysis (DLA) in order to find language features across millions of Facebook messages that distinguish demographic and psychological attributes. They show that their approach can yield additional insights (correlations between personality and behavior as manifest through language) and more information (as measured through predictive accuracy) than traditional a priori word-category approaches.

DataPlat'23: 2nd International Workshop on Data Platform Design, Management, and Optimization, March 28, 2023, Ioannina, Greece

mariella.bonomo@unipa.it (M. Bonomo);
simona.rombo@unipa.it (S. E. Rombo);
filippo.rotolo@community.unipa.it (F. Rotolo)

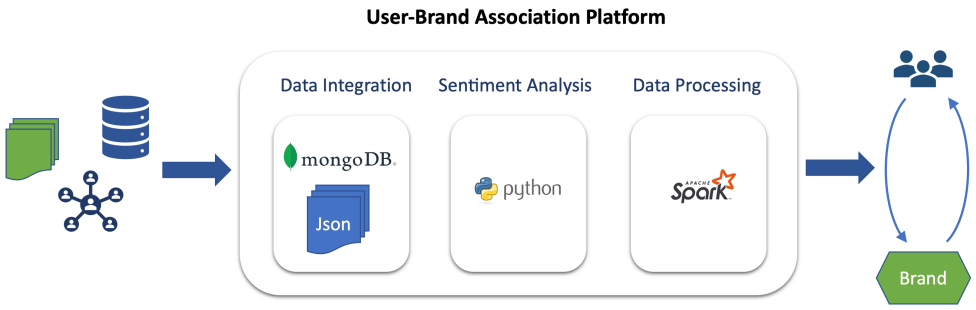

Figure 1: Design of the general big data platform for the prediction of user-brand associations.

The framework proposed in [9] relies on a semi- supervised topic model to construct a representation of an app's version as a set of latent topics from version metadata and textual descriptions. The authors discriminate the topics based on genre information and weight them on a per-user basis, in order to generate a version-sensitive ranked list of apps for a target user.

In [10] the authors propose a dynamic user and word embedding algorithm that can jointly and dynamically model user and word representations in the same semantic space. They consider the context of streams of documents in Twitter, and propose a scalable black-box variational inference algorithm to infer the dynamic embeddings of both users and words in streams. They also propose a streaming keyword diversification model to diversify top-K keywords for characterizing users' profiles over time.

Techniques applied to brand-affinity matching have been presented in [11, 12]. In particular, in [11] the authors present a profile-matching technique based on tree-representation of user profiles and apply it on Facebook ego-networks. In [12] a suitable combination of profile-matching and neighborhood analysis is used to identify the best k users for advertisements distribution.

3. Proposed Approach

In this section the approach proposed here is described in detail. In particular, it can be considered as the core of a more general platform, based on big data analytics as illustrated in Figure 1. The module *Sentiment Analysis* corresponds to the methodology described here.

Two main aims can be identified: (1) given a set of potential users, finding the best targets for an advertising campaign, and (2) given a new user in input, returning the best brand for which it can be a potential new customer. Sections 3.1 and 3.2 show these two different formulations of the presented approach.

3.1. Ranking Based on Sentiment Analysis

Let $U = \{U_1, U_2, \dots, U_n\}$ be a set of n users, representing possible customers of brands in the set $B = \{B_1, B_2, \dots, B_m\}$. The main aim here is to return a set $R = \{R_1, R_2, \dots, R_m\}$ of m ranks, one for each brand, each containing the list of users in U, ranked according to their potential "match" with the corresponding brand. This allows managers of advertising campaigns to select as targets only those users who may be potentially the most interested ones.

In order to understand to what extent each user matches with each brand, we assume that suitable tags may be associated with the brands. As an example, keywords may be extracted from textual exploration of their websites; also, such tags could be explicitly proposed by the campaign managers, according to the specific products object of the campaign. Let $T_1, T_2, \dots, T_m$ be the lists of tags associated to $B_1, \dots, B_m$, respectively. It is worth pointing out that different brands may share some common tags.

Let $u_i \in U\,(i = 1, \dots, n)$ be a user and $b_j \in B\,(j = 1, \dots, m)$ be a brand. The *Compatibility Index* between u_i and b_j is defined as:

$$I_C(i, j) = \frac{\sum_x M_x(i, j)}{|T_j|}$$

where $|T_j|$ is the number of tags associated to b_j, and $M_x(i, j)$ is the *match*, intended as a sort of liking rate, of the user u_i for the x-th tag of b_j $(x = 1, \dots, |T_j|)$.

In particular, the match $M_x(i, j)$ is obtained by sentiment analysis techniques [13]. The key aspect of sentiment analysis is to analyze the body of a text for understanding the opinion expressed inside it on some topics. Such an opinion, positive, negative or neutral, is usually referred to as *polarity* [14].

Polarity can be expressed as a numerical rating, representing a sort of sentiment score. There are different approaches to identify the sentiments expressed on topics. The main algorithms proposed for polarity computation can be distinguished in two main categories:

Table 1
Polarity values for the tags in the example.

	Ferrari			Oracle			Walmart			
	car	**red**	**expensive**	**technology**	**software**	**database**	**shopping**	**market**	**discount**	**usa**
Janny	0.0	0.5	0.0	0.0	0.0	0.0	1.0	1.0	0.5	1.0
Pedro	1.0	1.0	0.5	0.0	0.0	0.0	1.0	1.0	1.0	0.5
Andrea	0.0	1.0	0.0	1.0	0.5	1.0	0.0	0.0	0.5	1.0
Marian	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.5	0.0
William	1.0	0.5	1.0	0.0	0.5	0.0	1.0	0.0	0.5	0.5
Natan	0.5	0.5	1.0	0.0	1.0	0.0	1.0	0.0	0.5	1.0
Karmen	1.0	0.5	1.0	0.0	1.0	0.5	1.0	1.0	0.5	0.0
Sonia	0.0	0.5	0.5	0.5	0.5	0.0	1.0	1.0	1.0	1.0

- Supervised machine learning algorithms, trained to analyze any new text with a high degree of accuracy. This makes it possible to measure the sentiment, using for example support vector machines (SVM) [15].
- Unsupervised lexicon-based approaches using dictionaries of lexicon, (for example SentiWordNet [16], Flair [17], TextBlob [18], WordNet[19], Spacy-Textblob [20], AFINN [21], Sentistrength [22], Vader[6]).

Here, we have chosen the specific technique to apply for polarity computation by the experimental evaluation described in Section 4.1.

The output of the proposed approach is the set of ranks R, where each rank is a set of triplets $\langle u_i, b_j, I_C(i,j)\rangle$, sorted according to $I_C(i,j)$. A toy example is illustrated below.

Example 1

Let $U = \{Janny, Pedro, Andrea, Marian, William, Natan, Karmen, Sonia\}$ and $B = \{Ferrari, Oracle, Walmart\}$ be the users and brands stored in the input database, respectively. Ferrari has tags $T_F = \{car, red, expensive\}$, Oracle has $T_O = \{technology, software, database\}$, and Walmart has $T_W = \{shopping, market, discount, usa\}$, respectively. Table 1 shows the polarity values computed from the textual descriptions stored for users in the database (e.g., their tweets), while Table 2 shows the three ranks R_F, R_O and R_W obtained for the three brands Ferrari, Oracle and Walmart, respectively.

As the result of this toy example, Karmen, Pedro and William are the best targets for Ferrari; Marian and Andrea for Oracle; Sonia, Pedro, Janny, Karmen and Natan for Walmart.

3.2. Prediction Based on K-Nearest Neighbors

Let $C_1, C_2, \ldots, C_m$ be m classes, each associated to a brand in B, respectively. Objects in the classes are vectors of features, such that each vector represents an user and the features are the polarity values retrieved from the texts associated to the users in the input database, for the set of all tags associated to all brands. The K-Nearest Neighbors (KNN) classical approach can be then applied to predict the class label for each new potential customer, represented analogously by a features vector, by computing the (e.g., Euclidean) distance between this vector and those in the classes, and choosing as class label the most represented in the top K neighbors. The following example shows the KNN for the dataset in Example 1.

Table 2
Ranks obtained for Ferrari (R_F), Oracle (R_O) and Walmart (R_W), showing also the values of the Compatibility Index (I_C) between each user and brand.

Ferrari		Oracle		Walmart	
User	**I_C**	**User**	**I_C**	**User**	**I_C**
Karmen	0.83	Marian	1.00	Sonia	1.00
Pedro	0.83	Andrea	0.83	Pedro	0.87
William	0.83	Karmen	0.50	Janny	0.87
Natan	0.67	Natan	0.33	Karmen	0.62
Andrea	0.33	Sonia	0.33	Natan	0.62
Sonia	0.33	William	0.16	William	0.50
Janny	0.17	Pedro	0.00	Andrea	0.37
Marian	0.00	Janny	0.00	Marian	0.37

Example 2

Let Ferrari = {Karmen, Pedro}, Oracle = {Marian, Andrea} and Walmart = {Sonia, Janny} be three classes, such that the users are represented by the corresponding rows in Table 1. Suppose that Natan and William are the users for which the class labels are to be predicted, and $K = 3$. For Natan, the closest users are Pedro, Sonia and Janny, therefore it will be put in the class Walmart. William has Karmen, Pedro and Sonia as closest neighbors, therefore its predicted class label is Ferrari.

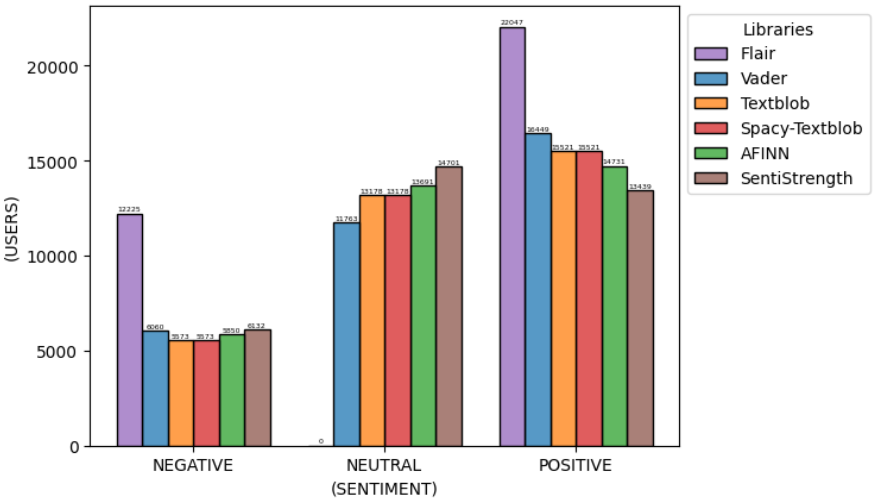

Figure 2: Comparison and test of Sentiment Analysis using different libraries: Flair, Vader, TextBlob, Spacy-Textblob, AFINN, Sentistrength.

4. Results

An important problem in the considered context is that it is often very difficult to find "true positive" and "true negative" samples in order to validate the proposed approaches. In particular, we have referred to the Twitter social network for the real data considered here, as explained in detail in Section 4.3. If one wants to validate the proposed approach, that is based on textual information stored in a database, by using independent information not related to the available textual one used for the classification, a simple method is to search for those users who are also "followers" of the considered brands. Indeed, to follow a brand, the user has made an action, and this can be considered as an explicit declaration of compatibility with the brand, that is not related to the user's textual information. Therefore, followers can be used as true positives. However, the same cannot be done for the true negative samples, due to the fact that negative relationships independent from the texts (tweets, in the considered case) are not available. For this reason, we have built a synthetic dataset, as explained in Section 4.2.

Before going through the description of results, we present in Section 4.1 a benchmarking analysis we have carried out, in order to decide the specific sentiment analysis algorithm to adopt for our experiments.

4.1. Benchmarking of Sentiment Analysis Libraries

Here we have referred to the unsupervised lexicon-based approaches for the polarity computation, due to the fact that they best fit with the considered context. In particular, Figure 2 shows the number of negative, neutral and positive polarity values returned by the considered algorithms (implemented by the corresponding Python libraries). It is evident that the results may be also very different for the evaluated sentiment analysis techniques, that use different dictionaries and approaches to calculate the polarity . For instance, the 'Flair' library tends to rate the tweets only positively or negatively, differently than the other ones.

In order to provide a comparison of the considered algorithms, a small set of 20 tweets for each tag has been manually verified to understand which approach has returned the right polarity value. Cumulatively, the best performing algorithm is Vader [6], which is also in accordance with the literature, since it uses a lexicon approach which has shown to be successful in the social media context.

4.2. Synthetic Dataset

Three classes have been built by choosing three tags for each, and by generating 15 users who like the corresponding brand and 15 who do not like it. The users have been obtained by generating sentences explicitly expressing positive (negative, respectively) opinions on that brand. Then, pairs of users and brands such that the user likes the brand, have been used as true positives, while pairs

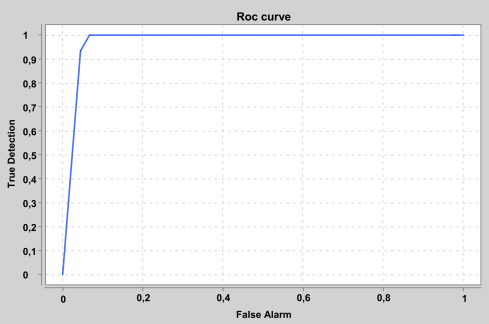

Figure 3: ROC curve for the synthetic dataset.

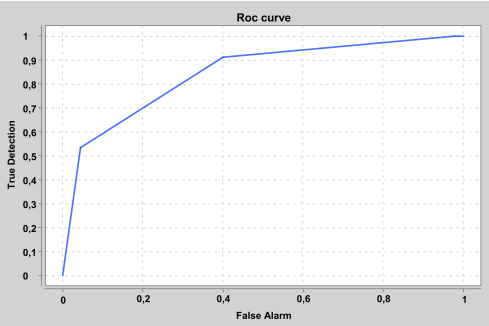

Figure 4: ROC curve for the synthetic dataset with noise.

with the user who does not like the brand as true negatives, respectively. Receiver Operating Characteristic (ROC) analysis has been used to verify the effectiveness of the approach, and Figure 3 shows the ROC curve obtained for this "perfect" dataset, for which the Area Under Curve (AUC) has resulted to be equal to 0.975.

Then, in order to verify also the robustness of the method, a small percentage of noise has been introduced by exchanging the 20% of sentences between the two types of users in the dataset. Figure 4 shows the ROC curve obtained in this case, with an AUC equal to 0.84.

4.3. Real Data Validation

The dataset coming from real data associated with input brands, and related targets, used for the validation is shown in Table 3. In particular, tags have been obtained from the keywords of the brands web pages.

As for the users' opinions, they have been generated as follows. The data of tweets for each user have been downloaded from Twitter by the Twitter APIs [7]. The Twitter APIs extract the data from Twitter accounts (in a certain date, time, number of followers or following, etc.). For the experiments described here, 129,274 tweets have been extracted from 1,046 users, chosen among the "followers" of the considered brands. Most tweets contain text and embed URLs, pictures, usernames, and emoticons. Therefore a pre-processing of tweets has been performed, such that tweets are filtered, and incomplete/inconsistent data eliminated. In more detail, each tweet has been suitably cleaned, by removing:

- URLs;
- tagged users names;
- special characters.

Moreover, all emoji symbols have been translated into text, and language contractions into their extended forms. Hashtags have been extended in sentences. In order to optimize the sentiment analysis process, the word lemmatization with part of speech has been applied for all extracted tweets. It is worth to point out that, in the pre-processing described above, stop words have not been removed, due to the fact that they can be important to establish the word polarity. Indeed, the sentiment analysis software libraries usually take them into account for this reason.

Tweets have been preprocessed also to eliminate duplicate tweets and retweets from the data, which led to a final sample of 12, 585 tweets and 498 users that can be considered significative (i.e., they have a sufficient number of tweets associated to perform the analysis). Sentiment analysis has been carried out using the Python VADER library. The VADER Sentiment Analyzer [23] combines qualitative and quantitative methods to produce a gold-standard sentiment lexicon and uses it to determine the polarity of tweets and to classify them according to multiclass sentiment analysis. This library uses the classification of the preprocessed tweets such that the polarity is considered:

- positive, with score's value 1;
- negative, with score's value 0;
- neutral, with score's value -1.

For the purposes of our research, the values returned by VADER have been normalized in the range $[0, 1]$.

Results obtained on real data shows the ability of the proposed approach in finding the best targets for advertising campaigns. Indeed, the 93.7% of effective followers have been correctly put in the first positions of the ranks, cumulatively. This shows that users who have provided preferences for some brands, as testified by the fact that they follow them, have scored high values of the Compatibility Index, with references to such brands, as expected.

The K-NN analysis has been performed by choosing three brands for which the tags are different and the resulting classes are thus well separated. The chosen brands are Adidas, Barilla and Microsoft. Only the followers of such brands have been put into the three classes. The user-vectors have been built from the polarities obtained, for each follower in the class, for the tags of the

corresponding brand. Then, the K-NN has been applied by leaving out the 20% of users for each class (test set) and keeping inside only the 80% of them (training set). The best resulting accuracy performed by the proposed approach is of 85.71% test users correctly classified, obtained for $K = 6$.

Table 3
Dataset of brands and related tags.

Brand	Tag
Alfaromeo	car, sporty, italian
Samsung	technology, computer, household appliances
Lavazza	coffe, italian, scent
Armani	clothing, fashion, italian
Adidas	clothing, sporty, fashion
Microsoft	technology, computer, office
Barilla	italian, pasta, home
Ferrero	chocolate, italian, baby
Delonghi	household appliances, home, coffe
Amarelli	chocolate, licorices, baby

5. Conclusion

We have presented a novel approach for the recommendation of new possible customers for existing brands. The approach is based on the assumption that textual information is stored in a database on both users and brands. Sentiment analysis techniques have been adopted to measure to what extent an user matches a brand, according to the retrieved users' opinions on different topics. Two different formulations of the proposed approach have been described, and the preliminary results obtained on both synthetic and real datasets are promising, looking at the high values of accuracy reached in both cases.

In the future, we plan to finalize the implementation of all modules of the general platform illustrated in Figure 1. Moreover, further work will regard the consideration of reciproque compatibility measures, in order to predict, for example, if users with high matches with a specific brand may also likely have high compatibility with other brands.

Acknowledgments

This research has been partially supported by the projects: "AMABILE - Amarelli BIg data and bLockchain Enterprise platform" (CUP: B76G20000880005) funded by the Italian Ministry of Economic Development, and "Big knowledge graphs modelling and analysis for problem solving in the web and biological contexts" (2022, CUP: E55F22000270001), funded by INDAM GNCS.

References

[1] R. Feldman, Techniques and applications for sentiment analysis, Communications of the ACM 56 (2013) 82–89.

[2] S. Bhatnagar, N. Choubey, Making sense of tweets using sentiment analysis on closely related topics, Social Network Analysis and Mining 11 (2021) 44.

[3] A. Mee, E. Homapour, F. Chiclana, O. Engel, Sentiment analysis using TF-IDF weighting of UK mps' tweets on brexit, Knowledge Based Systems 228 (2021) 107238.

[4] S. Barreto, R., J. Carvalho, A. Paes, A. Plastino, Sentiment analysis in tweets: an assessment study from classical to modern word representation models, Data Min. and Knowl. Disc. 37 (2023) 318–380.

[5] J. A. Hanley, et al., Receiver operating characteristic (roc) methodology: the state of the art, Crit Rev Diagn Imaging 29 (1989) 307–335.

[6] S. Elbagir, J. Yang, Analysis using natural language toolkit and vader sentiment, in: Proceedings of the International MultiConference of Engineers and Computer Scientists, 2019.

[7] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media?, in: Proceedings of the 19th international conference on World wide web, 2010, pp. 591–600.

[8] H. Schwartz, J. Eichstaedt, M. Kern, et al., Personality, gender, and age in the language of social media: The open-vocabulary approach, PLoS ONE 8 (2013) e73791.

[9] J. Lin, K. Sugiyama, M.-Y. Kan, T.-S. Chua, New and improved: Modeling versions to improve app recommendation, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14, ACM, 2014, pp. 647–656.

[10] S. Liang, X. Zhang, Z. Ren, E. Kanoulas, Dynamic embeddings for user profiling in twitter, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, 2018, pp. 1764–1773.

[11] M. Bonomo, G. Ciaccio, A. De Salve, S. E. Rombo, Customer recommendation based on profile matching and customized campaigns in on-line social networks, in: ASONAM'19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019, 2019, pp. 1155–1159.

[12] M. Bonomo, A. La Placa, S. E. Rombo, Identifying the k best targets for an advertisement campaign via online social networks, in: Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge

Management,Volume 1: KDIR, Budapest, Hungary, Nov. 2-4, 2020, 2020, pp. 193–201.
[13] R. Wagh, P. Punde, Survey on sentiment analysis using twitter dataset, in: Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 208–211.
[14] Z. Nanli, Z. Ping, L. Weiguo, C. Meng, Sentiment analysis: A literature review, in: 2012 International Symposium on Management of Technology (ISMOT), IEEE, 2012, pp. 572–576.
[15] T. Joachims, Making large-scale svm learning, Practical Advances in Kernel Methods-Support Vector Learning (1999).
[16] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010.
[17] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, Flair: An easy-to-use framework for state-of-the-art nlp, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations), 2019, pp. 54–59.
[18] J. P. Gujjar, H. P. Kumar, Sentiment analysis: Textblob for decision making, Int. J. Sci. Res. Eng. Trends 7 (2021) 1097–1099.
[19] C. Fellbaum, Wordnet, in: Theory and applications of ontology: computer applications, Springer, 2010, pp. 231–243.
[20] A. K. Singh, A. Verma, An efficient method for aspect based sentiment analysis using spacy and vader, in: 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 2021, pp. 130–135.
[21] F. Å. Nielsen, A new anew: Evaluation of a word list for sentiment analysis in microblogs, arXiv preprint arXiv:1103.2903 (2011).
[22] M. R. Islam, M. F. Zibran, Sentistrength-se: Exploiting domain specificity for improved sentiment analysis in software engineering text, Journal of Systems and Software 145 (2018) 125–146.
[23] A. Amin, I. Hossain, A. Akther, K. M. Alam, Bengali vader: A sentiment analysis approach using modified vader, in: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1–6.