Urdu Sentiment Analysis

Khairullah Khan[1], Atta Ur Rahman[3], Aurangzeb Khan[4], Ashraf Ullah Khan[6], Bibi Saqia[7]
Department of Computer Science
University of Science & Technology Bannu
Bannu, Pakistan

Wahab Khan[2]
Department of Computer Science and Software Engineering
International Islamic University, Islamabad

Asfandyar Khan[5]
Institute of Business and Management Sciences University of Agriculture Peshawar,
Pakistan

***Abstract*—Internet is the most significant source of getting up thoughts, surveys for a product, and reviews for any type of service or activity. A Bulky amount of reviews are produced on daily basis on the cyberspace about online products and objects. For example, many individuals share their remarks, reviews and feelings in their own language utilizing social media networks such as twitter and so on. Considering their colossal Quantity and size, it is exceedingly knotty to look at with and interpret specified surveys. Sentiment Analysis (SA) aims at extracting people's opinion, felling and thought from their reviews in social websites. SA has recently gained significant consideration, however the vast majority of the resources and frameworks constructed so far are tailored to English as well as English like Western languages. The requirement for designing frameworks for different dialects is expanding, particularly as blogging and micro-blogging sites are becoming popular. This paper presents a comprehensive review of approaches of Urdu sentiment analysis and outlines of relevant gaps in the literature.**

Keywords—Urdu; sentiment analysis; social media; survey

I. Introduction

In the near past about a decade, galore social networks are introduced while the already existed networks such as Twitter, Facebook and Instagram etc have blew up their presence on cyber space [1, 2]. The mentioned social networks have billions of users and yields a vast quantity of digital data, comprised of images, text, audio and video etc [2]. As indicated by [3], the evaluated measure of data on the web will be around 40 thousand Exabytes, or 40 trillion gigabytes, in 2020.

Online user comments are regularly utilized by individuals needing to know about different customer reviews about product or services of interest [4, 5]. These comments are frequently given in a free-text format by online business or social sites [4-6]. With the guide of this rich source of data individuals can certainly make a better decision [5, 6]. It is important to collect potential reviews in every field for good decision making and future prediction. Governments, Societies and organisations have an awesome enthusiasm for catching and collecting what individuals consider about a specific topic [1]. What the bigger portion of folks conceive Donald Trump, Facebook's procuring of WhatsApp and the fresh iPhone 10 are indispensable questions in the current age [1, 5].

Sentiment analysis can be used to solve such inquiries [1, 5]. Because SA is the process of investigating and extracting of individuals' remarks, audits and sentiments about a particular topic, an item, a news alert and a mobile application etc [1, 2, 5, 7]. The advantages of performing SA are countless [7]. It can help in measuring the public's opinion on controversial issues in a more precise, extensive and moderate form than open surveys [7]. It can likewise enable organizations to tailor their services, administration, product etc. to their clients' needs and in this way increment their benefits [7].

Abundant literature is available on strategies, difficulties, and utilization of SA for English language [1]. However, very little research has been led in SA of the Urdu language [8]. The SA designed for English language cannot be utilized for Urdu language having different script and morphological structure [8, 9]. Urdu has linguistic features that are diverse as well as contradicting to the English dialect [8, 9].

Like some other languages, Urdu online resource are in addition to getting popular with as individuals would like to partake impressions and convey opinions in their native languages [8, 9]. From Literature study it is clear that techniques employed for counterpart languages cannot be adopted for handling Urdu language issues [8, 9]. The main objective of this study is to report and investigate Urdu SA techniques and challenges.

The structure of the paper is organized as: section 2 describes SA in detail, in section 3 the recent trends and approaches of Urdu SA are provided, in section 4 literature review is provided, in section 5 Urdu SA challenges are described in detail, section 6 highlights SA techniques while in section 7 conclusion is provided.

II. Sentiment Analysis

Sentiments are vital to every human activity since they are critical influencers of our practice. At whatever point we have to draw a conclusion, we need to experience other people' feelings. In reality, organizations and businesses need to know about consumer and public opinion about their services and products. Individual clients additionally need to know the opinion of existing consumers of an item before obtaining it, and others' opinions about political candidates in election campaign. When an individual required opinion about someone, he inquired about him from his friends and family members. The executive members of an organization conduct external survey and organize polls in order to know public opinions to make a strategic decision.

Obtaining public and consumer opinions have long been an immense business itself for marketing, public relations, and political campaign.

With the development of web-based social networking (e.g. Twitter, Facebook, and YouTube etc.) on the Web are progressively utilizing the contents available on the social media to make a suitable strategic decision. These days, if one needs to purchase a product, one is no more restricted to review people opinion on the Web about that item. For an organization, it is not obligatory to carry on surveys, public opinion poll, and center groupings called for to get together masses' impression, because there is a lot of such kind of data openly accessible on the Web.

However, coming up and checking out opinion sources on the WWW and rectifying the data arrested in them persists an imposing errand in view of the proliferation of a different site. Each site commonly contains a large volume of sentiment message that isn't generally effortlessly decrypted in long in long sites and discussion postings. A moderate human reader will experience issues recognizing authoritative sites and pulling out and summing up the opinions in them. Automated SA frameworks are thus needed. SA application have spread to practically every possible domain, such as customer services, consumer product, financial and health care services to political campaign etc [10].

III. Urdu Sentiment Analysis

This section describes the overview Urdu SA. Urdu language has unique feature and therefore some extra steps are needed for SA process. For example, Urdu script is written from right-to-left. The shape of alpha bits changes with change of its position in words. Urdu has different stop words. Some common grammatical mistakes are raised while dealing with stop words. For example کہ is confused with کے , پہ with پے and other rhyming words but having different meanings.

In Urdu writings the usage of space is not consistent and usually leads to either space insertion or space omission problem in words. The Space omission problem e.g. the Urdu word " انکا " which is in reality a compounding of cardinal words, but the system processes it as an individual word. The space Insertion problem e.g. the word "عقلمند " (Aqalmand, Intelligent) s in reality a single word but when tokenized , will be handled as compound word i.e "عقل" and " مند" which is addressed by a two-stage system.

The literature shows various SA techniques that have some common steps. The most commonly used steps are given in figure 1.

The first step is pre-processing. In this step a series of sub steps are carried out e.g. Noise Removal, Sentence Boundary Detection, Words Tokenization, and in some cases Part of Speech Tagging. Hence the related problems are needed to be handled in pre-processing step.

The next commonly used step is polarity identification i.e. for a word, sentence or text it to determine that it is positive, negative or neutral.

In order to prepare different models or algorithm successfully, it's essential to provide right training data to the models and furthermore the data must be sufficiently vast to prepare the model effectively [8]. The Urdu dataset can be collected from the commonly used Urdu News portal such as BBC Urdu, AAJ News, Abb Takk News, ARY News, *Geo* News, Samaa News, Dunya News, Dawn News, Express News, jang News etc

In the next phase tokenization of Urdu text is performed. Tokenization is the way toward separating the given content into units called tokens [8, 9]. The tokens might be words, numbers or punctuation marks. Tokenization does this assignment by finding word boundaries [8, 9]. When tokens are created from the sentences then it is passed to the polarity identification stage [8, 9]. Polarity of each word is determined by comparing it with sentiment lexicon [8]. Polarities of each word are determined as: Positive=1, Negative= -1, and Neutral=0 [8, 9].

When single polarities are assigned to each word then the combine polarity of the sentence is calculated [8, 9]. For example if a specific sentence has two positive words and one negative word, overall polarity would be determined as + 1(+2-1), consequently declaring it as a positive comment [8, 9]. The output demonstrates that the remark/feeling of the reviewer had a positive, negative or neutral conclusion towards the news, items or products [8, 9].

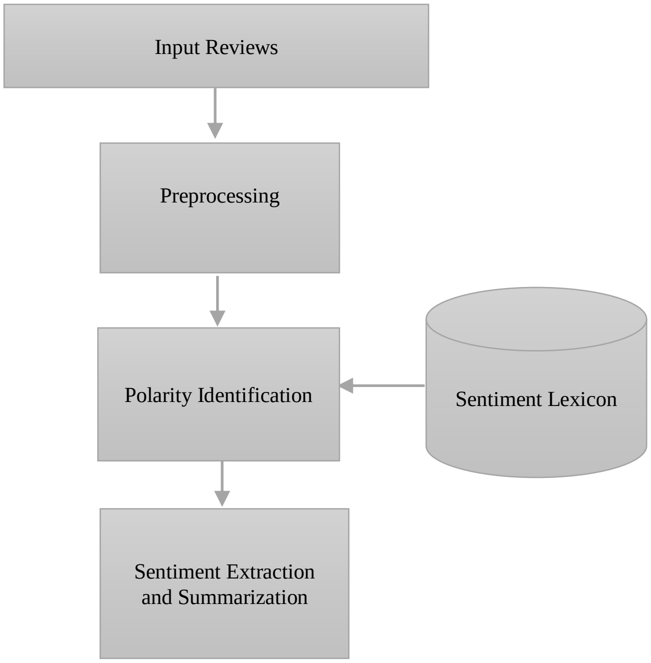

Fig. 1. Common Steps of Sentiment Analysis.

TABLE I. SUMMARY OF EXITING WORK ON URDU SENTIMENT ANALYSIS

Authors	Year	Task	Model/Approch	Polarity	Data scope	Data set/source	Language
Afraz Z. Syed	2010	Lexicon Based SA	Classification	Pos/Neg	Urdu Web Forums	Movies and Products reviews	Urdu
Afraz Z. Syed	2011	Adjectival Phrases as the Sentiment Carriers	Classification	Pos/Neg	Urdu Web Forums	movies and electronic appliances	Urdu
Faiza Hahim	2011	Lexicon Based SA	Classification	Pos/Neg/Neutral	Urdu News Headlines	Product and Movie reviews	Urdu
Smruthi Mukund	2011	identify Opinion Entities	SVM/Kernels Method	N/A	Urdu News Headlines	BBC Urdu News Portal	Urdu
Smruthi Mukund	2012	Analyzing Urdu Social Media for Sentiments	SVM	Pos/Neg	Newswire data	cricket and movies	Urdu
Syed Afraz Z	2014	Identification and Extraction of Appraisal Expressions	Classification	Pos/Neg	Urdu Web Forums	Movies and Products reviews	Urdu
Misbah Daud	2015	Opinion Mining System	Machine Learning	Pos/Neg/Neutral	Roman Urdu	1620 comments	Roman Urdu
S. Abbas Ali	2016	Salience Analysis of NEWS Corpus	Heuristic Approach	Pos/Neg	whole News except the heading	Urdu News Corpus	Urdu
Muhammad Bilal	2016	Sentiment classification	Classification: Models used are Naïve Bay, Decision Tree (DT) and KNN	Pos/Neg	Roman-Urdu and English	The model performance was evaluated on dataset of 150 positive and 150 negative reviews	Roman Urdu
A. Nazir	2017	Opinion Extraction	lexicon-based approach	Pos/Neg	Urdu Web Forums	100,000-tagged words downloaded from http://www.cle.org.pk	Urdu

IV. LITERATURE REVIEW

There are galore practical application and sweetening on SA approaches that were advised over the latest couple of years. This study expects to give a more intensive look on these improvements and to summarize and classify a few articles displayed in this field as indicated by the different SA systems. The author has collected different articles which represent critical upgrades to the SA systems utilized for Urdu language. This paper covers a wide variety of SA fields published in the last few years for Urdu language processing. They are arranged by the objective of the article showing the algorithm, data sets and information utilized as a part of their work. This study can be helpful for new comer scientists in this field as it covers the most well-known SA procedures and applications in a single research paper. It talks about additionally new related fields in SA which have pulled in the analysts of late and their comparing articles.

Websites such as IEEE Explore, Springer Linker, Science Direct, ACM Portal and Googol Search Engine were utilized as a base for this exploration [11]. Diverse keyword terms such as ('Urdu ‘and ' Sentiment Analysis') and ('Urdu' and 'opinion mining') etc were searched for exploration. The articles displayed in this survey are outlined in Table 1.

V. CHALLENGES OF SENTIMENT ANALYSIS IN URDU LANGUAGE

Datasets: There are rear datasets and corpora available to apply for SA in Urdu language. Data are generally collected from social networks and online forums, and newspapers.

Lexicon: In lexicon based SA we require a sentiment annotated lexicon of Urdu words constructed from a huge amount of text. The most suitable and easy way of text collection is online resources such as web blogs, social media sites, online news or electronic journals etc. Although Urdu is a very rich language its available resources on the internet are limited [12]. Furthermore, most of the data are Urdu websites are available in graphics/image format and hence are not easily retrievable. There are few publicly available corpus for Urdu sentiments and few lexicons have been created so far [9] and most of them are not openly available.

Opinion Target and Opinion Words Detection

For opinion mining and SA identification of opinion target is an important task […]. Noun has been most popularly employed for opinion target detection while adjectives have been employed as opinion words. In Urdu language we have problem e.g in the sentence

اسلام محبت اور امن کا درس دیتا ہے

The words اسلام, محبت and امن are nouns where اسلام is opinion target about which positive opinion is represented through the words محبت and امن. Although adjective is not available in this sentence it is still opinionated. Therefore, the opinion mining system must be able to classify nouns as opinion target and opinion words based on the sentence structure and opinion lexicon [..].

Another serious issue in Urdu language is that in day to day activities, online blogs and even in newspaper we observe English words while talking in Urdu. Similarly, Roman Urdu is most popularly used in most of the online platforms such Facebook, Twitter etc. In Such case traditional system fails to detect opinion words opinion targets from Urdu text. Hence the system should have additional preprocessing step such as to detect, clean and convert in normal text.

Detecting opinion spam

The issue of opinion spam is also observed in Urdu SA. Opinion spam is fake sentiment utilized for misguiding the users. This is typically done in organizations to advance or business to promote.

Resolving co-reference

Co-reference resolution also observed in Urdu SA, it is indistinct when the sentiments refer to multiple opinions.

Feature extraction

Urdu text is usually unstructured which make morphological analysers and POS taggers extremely difficult for Urdu language processing.

Segmentation: Segmentation issue can be further classified as a) Space-inclusion, b) Space-deletion issues. For example a single word can have a space in it, such as, " حوب صورت "(khoob surat, beautiful). On the other hands, space between two distinctive words can be deleted such as, "دستگیر"(dastgeer, benefactor) [9].

VI. Sentiment Classification Techniques

The core sentiment classification approaches can be placed in three broad categories e.g. machine learning approach, lexicon based approach and hybrid approaches See Figure 2 [13].

The term "Machine Learning" is identified with enabling computer to learn training it first on pre labelled data that enables it to arrive at a prediction about test data that it may be given over later time [13].

In various research analysis the data sets are annotated manually, while some annotate it automatically by utilizing online lexicons such as SentiWordNet etc [13]. There are not many resources available for Urdu SA. Subsequently, the majority of the past research annotated the data manually. In SA, supervised ML is performed by giving the computer a set of phrases (features) and their polarity, giving the computer the capacity to predict the polarity of the unseen text [13, 14]. Unsupervised ML is utilized when it is hard to discover a labeled feature.

Supervised ML classifier can be divided into four principle classes, such as decision tree, linear, rule base and probabilistic classifiers [13]. Where linear classifier can be further classified into Support Vector Machine (SVM) and Neural Network (NN). Probabilistic classifier utilizes the probability of specific term/phrase to predict the polarity of unseen text. It can be further divided into Naive Bayes (NB), Bayesian Networks (BN) and Maximum Entropy (ME) classifier [13].

The lexicon based strategy is typically utilized when the data are unlabelled [1]. Lexicon are utilized to label the data and to predict the polarity of each word. In the Urdu dialect, only a few lexicon are available [1]. A few researcher have made Urdu lexicon, however, the greater part of these lexicons are not openly accessible [1].

The hybrid approach utilizes both lexicon and machine learning-based techniques [1]. This approach is more predominant in the current literature and have a higher efficiency as compared to lexicon based and machine learning techniques alone [1]. The lexicon scores are typically utilized as a features in the classifier [1].

In dictionary based approach a set of sentiment words is collected manually with their orientation [1]. This set is developed by searching in the WordNet or thesaurus for their equivalent words and antonyms. The recently discovered words are added to the seed list then the following emphasis begins [1]. The iterative process stops when no new words are found. After the access is finished, manual investigation can be performed to evacuate or correct errors [1].

The Corpus-based approach tackles the issue of finding opinion words with context specific orientation [1]. Its strategies rely upon syntactic examples or examples that happen together along with a seed list of opinion words to find other opinion words in a vast corpus [1]. Corpus based approach can be further classified into Statistical and Semantic approach. In a statistical approach the polarity of a word can be identified by finding the occurrence frequencies of the word in a corpus. Semantic technique gives similar sentiment values to semantically close words [1]. For example WordNet gives various types of semantic connections between words used to find sentiment polarities [1].

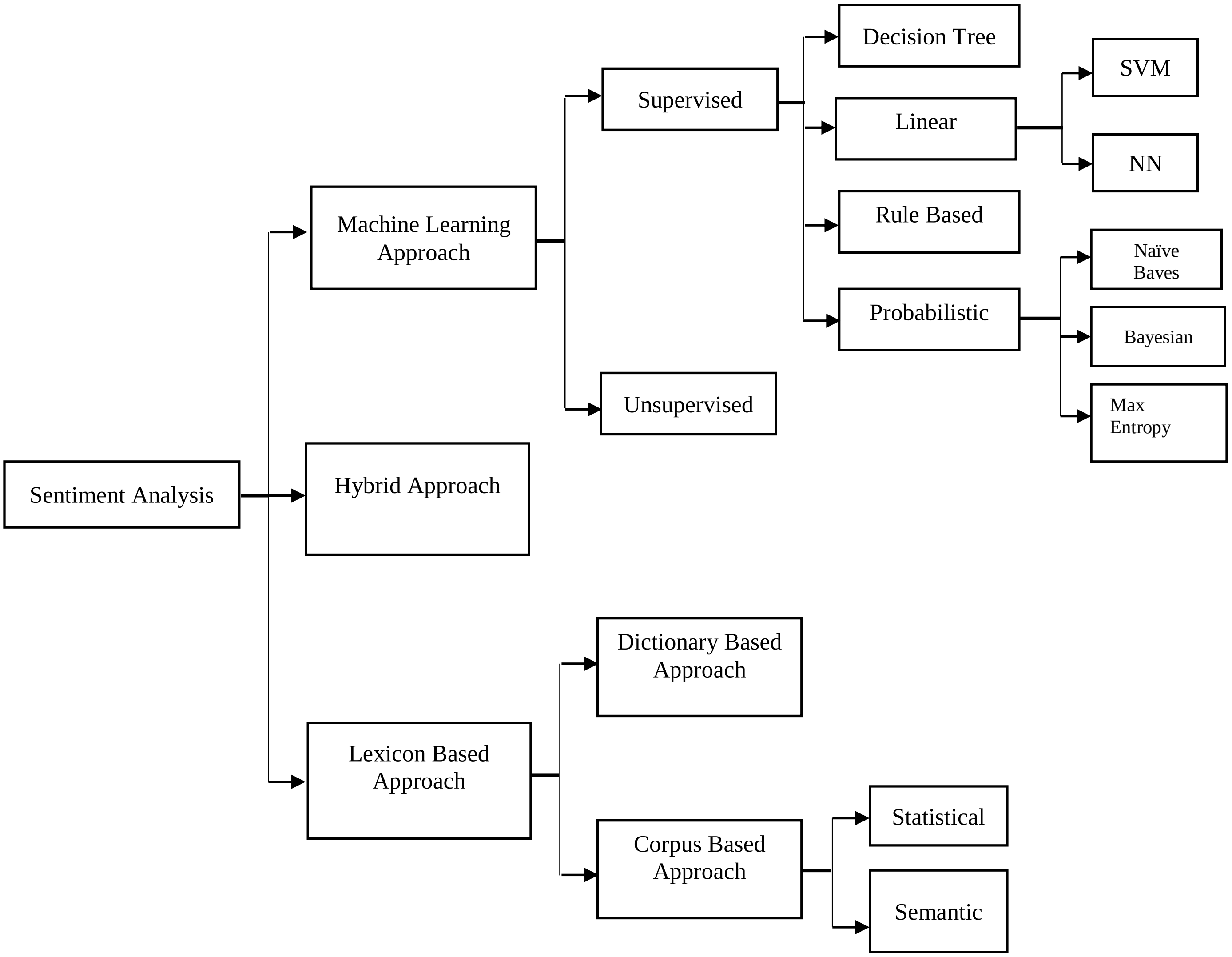

Fig. 2. Sentiment Classification Techniques.

VII. Conclusion

It is evident from this study that a lot of potential strategies and methodologies are available but still little work is done on Urdu sentiments analysis. Table 1 summarizes all related articles to Urdu SA published up to date. Masses of clients share their opinions via web-based networking media, making it a significant platform for tracking and exploring public opinions. Online networking is one of the greatest stages where huge texts are published each day which makes it a perfect source for catching people opinions on different topics such as products, services and celebrities etc. The primary objective of this paper is to give an outline of most recent updates in SA and classification techniques utilized for Urdu language. Various improvements are possible in the field of Urdu SA, such as utilizing different data sources from multiple Urdu web forums. Another critical expansion is to exploit effective techniques that increase the use of online market and social media to extract sentiments from user reviews or comments for the purpose of improving products and services.

References

[1] M. El-Masri, N. Altrabsheh, and H. Mansour, "Successes and challenges of Arabic sentiment analysis research: a literature review," Social Network Analysis and Mining, vol. 7, p. 54, 2017.

[2] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic Language Sentiment Analysis on Health Services," arXiv preprint arXiv:1702.03197, 2017.

[3] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," IDC iView: IDC Analyze the future, vol. 2007, pp. 1-16, 2012.

[4] T. Shivaprasad and J. Shetty, "Sentiment analysis of product reviews: A review," in Inventive Communication and Computational Technologies (ICICCT), 2017 International Conference on, 2017, pp. 298-301.

[5] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghassem-Aghaee, "A framework for sentiment analysis in persian," Open Transactions on Information Processing, vol. 1, pp. 1-14, 2014.

[6] M. Korayem, D. J. Crandall, and M. Abdul-Mageed, "Subjectivity and Sentiment Analysis of Arabic: A Survey," in AMLTA, 2012, pp. 128-139.

[7] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on, 2013, pp. 1-6.

[8] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Lexicon based sentiment analysis of Urdu text using SentiUnits," in Mexican International Conference on Artificial Intelligence, 2010, pp. 32-43.

[9] Z. U. Rehman and I. S. Bajwa, "Lexicon-based sentiment analysis for Urdu language," in Innovative Computing Technology (INTECH), 2016 Sixth International Conference on, 2016, pp. 497-501.

[10] B. Liu, "Sentiment analysis and opinion mining," Synthesis lectures on human language technologies, vol. 5, pp. 1-167, 2012.

[11] K. Khan, B. B. Baharudin, and A. Khan, "Mining opinion from text documents: A survey," in Digital Ecosystems and Technologies, 2009. DEST'09. 3rd IEEE International Conference on, 2009, pp. 217-222.

[12] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "Urdu Named Entity Dataset for urdu Named Enity Recognition Task," in 6th International Conference on Language & Technology 2016, pp. 51-55.

[13] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, pp. 1093-1113, 2014.

[14] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," Kuwait journal of Science, vol. 43, pp. 66-84, 2016.