OPEN ACCESS JOURNALS

Indiana Journal of Multidisciplinary Research

Abbriviate Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.12679682

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-04|Issue-03|2024

Social Media Assisting Platform Using Sentiment Analysis

Anthony Shashanth[1], Praveen Patil[2], Surya K M[*3], Swati V Bhat[4], Kiran Kumar K[5]

[1,2,3,4] Information Science and Engineering, RV Institute of Technology and Management, Bengaluru, Karnataka, India

Article History
Received: 20.05.2024
Accepted: 05.06.2024
Published: 30.06.2024

Citation
Shashanth, A., Patil, P., Surya, K. M., Bhat, S. V., & Kumar, K. K. (2024). Social Media Assisting Platform Using Sentiment Analysis. *Indiana Journal of Multidisciplinary Research*, *4*(3), 179-186.

Abstract: A increasing trend in the automated society we live in today is to employ machine learning to accurately analyze social media conversations. There is no denying social media's pervasive influence on people's opinions and thought processes. The proliferation of social media and the Internet has led to the generation of large amounts of data, which can be exploited. Sentiment analysis on social media posts is one such application that can offer insightful data for corporate intelligence, social mobility, public opinion polling, and Internet of Things (IoT) motivational gadgets. Emotional intelligence (ER) and applied emotional research are the main topics of this paper. The goal of our suggested framework is to represent people's actual thoughts and feelings. We present a novel dictionary- based method that, instead of ignoring or generalizing, handles long words as unique characteristics. Specific syntactic rules are used to determine sensitivity scores for long words. The overall number of sensitivities is then calculated by adding these scores. Informal discussions from social networking sites like Facebook, Twitter, and private chats are included in our data structure. With F-measure rates ranging from 81% to 96% for all data sets, the comparison study demonstrates that our suggested technique, which breaks long words, performs better than the conventional algorithm.

Keywords: Natural language processing, social media, sentiment analysis, emotional recognition, and extended words.

INTRODUCTION

People are drawn to other people who have similar values, opinions, and hobbies. According to studies, people find it easy to associate with people who support their aims and who hold similar beliefs.

This tendency to create like-minded people is deeply ingrained in human nature. Communities are clustered, and modularity is a key factor in determining their structure. A closer examination of the characteristics of these groups allows the unique characteristics of each group of like- minded individuals to be identified. Essentially, a common bond within a group of people implies a common set of ideas and goals.

Types of social media models, each serving different purposes, as shown in Fig.1 Table 1 provides an example of the categories of these social media models with websites and applications under each category social media serve as channels for sharing information, ideas, and presentation through virtual networks Popular social networking sites like Instagram, WhatsApp, Facebook, and Twitter have become important platforms for people to express themselves through text posts, status updates, and other types of content like text, images, videos, and audio. In order to create a more participatory environment, these platforms enable users to exchange, collaborate, discuss, and edit user-generate material.

Due to the abundance of information on social media, traditional surveys, opinion polls, and focus groups are becoming more and more necessary to determine public attitude. Opinions expressed on social media have a greater influence on politics, the general mood, and business these days [1], [2].

But this raw data is often available in unstructured or semi-structured formats, making simple data points like likes, shares and comments insufficient for meaningful analysis.

It is vital to process this raw or semi-structured data in order to get useful insights from it. This topic includes removing idioms, lengthy words, syntax, semantics, and other pertinent [3],[4].

***Corresponding Author:** Surya K M

Table1: Sort of Services Associated with Social Media

Types	Services	Websites
Entertainment Models	Game Play and Sharing Virtual Worlds	Candy Crush, Kongregated, Miniclip, Anipang, The Sims, Second Life
Collaboration Models	Q & A, Community Wikis Reviews and Opinions Social Bookmarking	Askville, Yahoo!, Answers, Spring.me Twiki, Pbworks, Evemote, Eopinions, Kindle, Amazon, Delicious, Readwrite, Scoopit, Digg, Diigo
Sharing Models	Magnines, Files, Books, Dcouments Music and Audio Livestreaming Photo Video	Scribd, Google Docs, Issuu, 4shared Last.fm, Tunes, Soundcloud Justin.tv, Ustram.tv, Flickr, Instagram, Vine, YouTube, Vimeo
Communication Model	Video-Conferencing Instant Messaging Evrnt Networking Social Networking Microblogging	Skype, Google Hangout, Viber, WhatsApp, Kakao Talk, Line Meetup.com, Upcoming Facebook, Google+, LinkedIn, Ning, Myspace, Twitter, Tumblr, Me2day

Problem Description

The topic of long words—words with repeated letters included in common dictionary words—is discussed in this essay. For example, if "happy" is a traditional dictionary word, then its tall equivalent must be "happyyy." Notably, as Figure 2 illustrates, people are increasingly choosing to use such long terms to convey their feelings, particularly on social media. These words therefore have a lot of meaning in the context of social networks. These terms were frequently dropped from sentences in the past because they were thought to be meaningless.

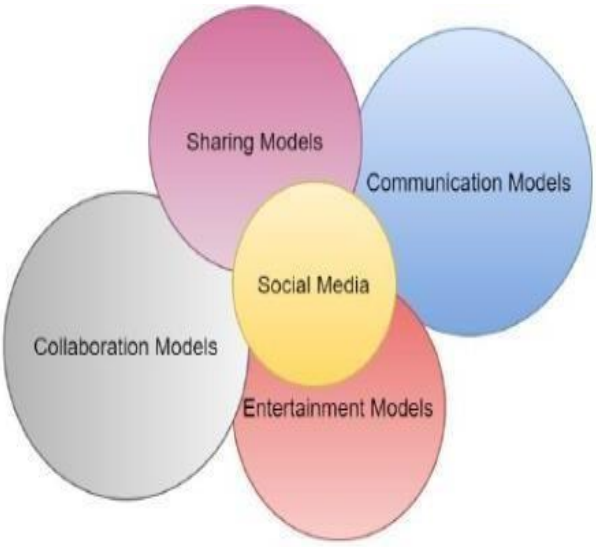

Figure 1: The Various Models of Social Media.

It was eventually discovered that this procedure altered the message's intended meaning. Because they can improve sentence context and express nuances of meaning, academics have argued that normalizing rather than eliminating these longer words. However, current research suggests that routinely using long words may not be an effective technique for dealing with long words.

General terms often ignore or minimize the emotional strength of these words. Individuals often use long words to express strong emotions such as extreme happiness (e.g.,"Awesomeeee"), sadness (e.g., "badddd"), or anger (e.g., "no wayyy") Generalization is so tiring translating these words into their dictionary equivalents (e.g. "Awesomeeee" becomes "Awesome", "badddd" becomes "bad", and "no wayyy" becomes "no way"), thus giving a true feeling or pleasure the amount the person expresses decreases.

Figure 2: Example of YouTube.

Goals

The following are the goals of this paper.

- To examine current sensitivity analysis programs that oversimplify capitalized terms.
- Introduce a new system for sentiment analysis based on dictionaries that uses capitalized terms.
- To use conversational data collected at random for research reasons from friend groups on Facebook, Twitter, YouTube, and in private discussions.
- To assess the suggested system's performance using metrics like F-measure, precision, and recall.
- To compare the suggested approach with the conventional algorithm which disregards lengthy terms in all data types in a comparative manner. The purpose of this comparison is to confirm the efficacy and precision of the suggested approach.

***Corresponding Author:** Surya K M

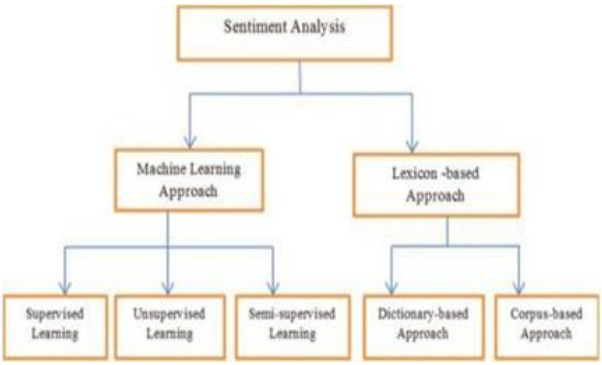

Figure 3: Framework of Paper Classification

The structure of the paper

The structure of the paper is as follows: Related work on sensitivity analysis is described in Section II. Section III describes in detail the materials and methods used. Section IV illustrates the proposed framework with examples to clarify the effect of long words. Section V presents the experimental findings of the suggested system. The study's perspective and conclusion are finally presented in Section VI.

RELATED STUDY

Emotion analysis also known as attention mining or emotion polarity assessment involves extracting analysis from text based on text structure and place of use This analysis divides different methods into different groups based on different perspectives of emotion on research programs. These include classical methods, transfer learning (TL) methods, and deep learning (DL) methods, which aim to explore and explore recent relevant literature in the field.

Machine learning approach

Intelligence (AI) incorporates machine learning (ML), which aims to improve decision-making and prediction by uncovering important patterns in data [5]. Compared to other methods, sentiment analysis (SA) has significant advantages over ML due to its ability to handle large amounts of data from the Internet and automation ML algorithms are designed to identify patterns in data and generate these patterns associated with specific pattern patterns in the data. ML is classified as supervised learning (SL) and unsupervised learning (UL).

Supervised learning:

In supervised learning, there are two primary techniques: regression and classification. Output variables used in classification are categorical and often comprise two or more categories. Regression, on the other hand, looks at the relationship between two or more variables, determining how changes in one are connected to changes in another. [6]

Unsupervised learning:

Unsupervised learning focuses on the machine's ability to recognize patterns and respond appropriately without explicit instructions. UL can be further divided into Clustering and Association. Clustering involves grouping data based on behavioural similarity, while association uses ML-based rules to reveal significant relationships between variables in large data sets.

Lexion-Based Approach

The dictionary-based method is an approach to perceptual analysis based on assumptions about the percepts derived from individual words or groups of words [7] Commonly referred to as keyword- based methods, this approach uses text a dictionary containing words related to emotions or at uses at least words or phrases associated with a particular emotion or becomes a list This dictionary helps to identify the content of the message, and the emotions and feelings of the words a contains express.

Dictionary-based method

Also known as the rule-based method, the dictionary-based method is a sensitivity analysis technique that makes use of a pre-established dictionary or lexicon to assign a sensitivity score [8] A sensitivity score is assigned to a word each in a text based on it corresponds to the entries in the dictionary. The overall sensitivity score of the sentence is then computed by summing or averaging the sensitivity scores of the connecting words. This method is useful for sensitivity analysis as it allows for rapid and efficient analysis of large amounts of text.

Corpus Based Methods

The corpus-based approach looks for patterns and regularities in language use by analysing multiple collections of texts, called corpora. In contrast to the dictionary- based approach, that approach this is based on preconceived feelings or assumptions about language function Tests the application of language incorrectly. Instead, it takes a logical and data-driven approach to language analysis. By analysing corpora, this approach seeks a corpus-based approach that can reveal common words and phrases, as well as complex linguistic structures such as collocation, formulaic order, and grammatical structure.

MATERIALS AND METHODS

Contents:

This section describes the methodology used. Includes discussion of environmental design, data collection methods, and data processing techniques. Additionally, it provides details of the sensitivity analysis process that incorporates capitalized capitalization.

***Corresponding Author:** Surya K M

Table 2: List of Sentiment Analysis Task

Sentiment Analysis Tasks	
Tasks	-
Spam detection	-
Sentiment Lexicon generation	-
Word sense disambiguation	-
Sarcasm detection	-
Time, Entity and opinion holder extraction	-
Sentiment search and retrieval	-
Polarity classification	Cross Language sentiment classification Sentiment rating prediction Cross domain sentiment classification Contextual polarity disambiguation Aspect-based sentiment classification
Sentiment summarization	-
Subjectivity classification	-

Environmental Activities

Python is widely recognized as one of the leading programming languages in the fields of natural language processing (NLP), machine learning (ML), and data analysis. Its extensive library ecosystem offers many NLP and ML techniques applied to a variety of problems. Python was selected as the study's programming language because to its extensive library support and intuitive user interface. Notably, the Python module NLTK makes it easier to work with data pertaining to human language. Together with useful modules like matplotlib, NumPy, Panda, and Seaborn, it also contains a tasty soup. Additionally, the system now includes new extraction techniques.

Data Sets

The data used in this study included informal conversations between groups of friends on Facebook, Tweets, and chat forums. It is noteworthy that a large proportion of young people express their emotions informally through texting, which is often reflected in the use of long words Data files are formatted in comma separated values (CSV) format for ease of handling in the Python environment. The intensification criteria used in the suggested algorithm has been created by three linguists, resulting in a kappa agreement of 0.68.

Data Preparation

Simple Python programming has been created to remove unnecessary elements to organize the necessary data. Experimental development of perceptual scores.

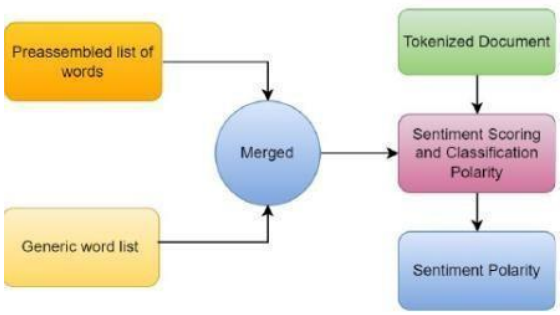

Fig.4 Different Datasets

THE PROPOSED METHOD

The proposed method includes the following steps, as seen in Figure 5, with the goal of processing unprocessed data from different sources, initially in an unstructured order:

Phase1: Tokenization is the initial stage.

The data is tokenized in this stage, which separates it into tokens using delimiters like spaces, commas, hashtags (#), and "@" symbols.

Phase 2: Stop the password search

Tokens developed in the tokenization phase tend to stop word removal, where common words are rejected for no reason, such as "the" in addition to emojis are also removed from the text because they do not contribute to in sensitivity analysis.

Phase 3: Normalization and senti-score generation

The remaining tokens go through this phase for normalization, where the tokens are standardized. Meanwhile, the fourth tranche of tokens is also processed. Each content word has a Senti score in Senti WordNet, with the Senti score of the regularized word subtracted in this step.

Phase 4: The fourth factor is tree removal and the severity of the tree score

Separate tables are built for every token created in the preceding section in parallel. The tokens' distinct characteristics are listed in the first column of these tables. The normalized word from the normalization module appears in the second column, and the number of times each character in the tree spelled a longer word from a previous step—which is kept in the first column—is counted in the third column. Subsequently,every cell is added together in the third phase to determine a tree score for every token. To get a robust tree score, multiply the tree score by 0.01.

Phase V: Aggregate acute senti-score

After the fifth step, the outcome of the centi-score generation and the weighted tree scores are

***Corresponding Author:** Surya K M

combined to calculate the weighted centi-score. These scores contribute to the final prediction, where they are added to the terms to determine the evaluation score. Analytical classification thresholds have been established, with scores below -0.05 classified as negative, scores greater than 0.05 as positive, and median scores as neutral.

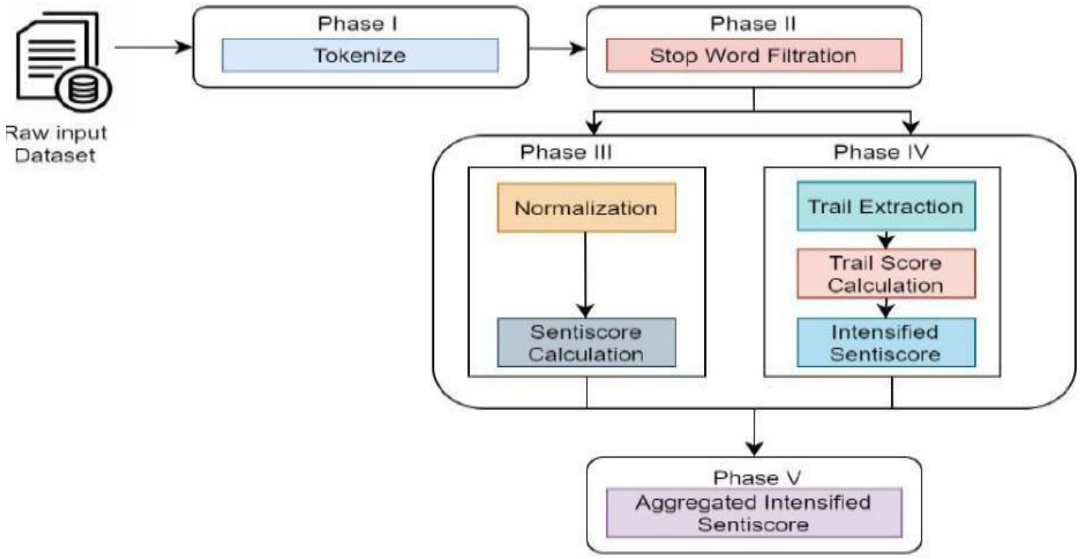

Figure 5: Proposed System for Detecting Lengthened Words in Sentiment Analysis

EXAMPLE

After taking the input, the proposed system starts working by going through the different phases.
Raw data: I am feeling veryyyyhapppyyyy.

Phase I: Tokeniser - During this phase, the words are extracted.
‘‘I’’, ‘‘am’’, ‘‘feeling’’, ‘‘veryyyy’’, ‘‘happpyyyy’’.

Phase II: Stop words are removed in this phase.
‘‘feeling’’,‘‘veryyyy’’,‘‘happpyyyy’’

Phase III: This phase deals with the normalization of the words using standard English WordNet. Al though here we normalize the word but retaining the original input is also required. For that a separate table is maintained to keep the flavor of lengthened alive. This further used in next phase for caring out the intensification.

Normalized words	Retaining original words
1. Feel	variable1=feeling
2. Very	variable2=veryyyy
3. Happy	variable3=happpyyyy

Phase IV: This phase is the backbone of the proposed system. Here the actual intensified senti-score is being calculated.

	Unique Character	Dictionary Count	Tail
Let Senti-score for “feel” is x1	f	1	0
	e	2	0
	l	1	0
	i	1	0
	n	1	0
	g	1	0

Table for variable 1= feeling

	Unique Character	Dictionary Count	Tail
Let Senti-score for “very” is x2	v	1	0
	e	1	0
	r	1	0
	y	1	3

Table for variable 2= veryyyy

	Unique Character	Dictionary Count	Tail
Let Senti-score for “happy” is x3	h	1	0
	a	1	0
	p	2	1
	y	1	3

Table for variable 3= happpyyyy

Intensified Tail Score for variable 1= 0*0.01
Intensified Tail Score for variable 2= 3*0.01
Intensified Tail Score for variable 3= 4*0.01

Intensified Senti-score for “feeling” =x1 + Intensified Tail Score for variable 1
Intensified Senti-score for “veryyyy” =x2 + Intensified Tail Score for variable 2
Intensified Senti-score for “happyyyy” =x3 + Intensified Tail Score for variable 3

***Corresponding Author:** Surya K M

Phase V: Finally, aggregation of the intensified score of each token of a document give rise to the actual intensified senti-score of a document.

EXPERIMENTAL RESULTS

The results of the test at the word representation are obtained through tokenization.

Performance Metrices

Precision, recall, and precision are used for analysis. In a retrieval system, accuracy refers to the proportion of relevant documents in the retrieved system.

Precision of an information retrieval system is defined as the proportion of the relevant documents in the retrieved set.

P=True Positive/(True Positive +False Positive)

Recall is defined as the proportion of the relevant documents in the collection that has actually been retrieved.

R=True Positive/(True Positive +False Negative)

F-measure is defined as the harmonic mean of Precision and Recall.

$F-measure = 2*PR/(P+R)$

The outcome of the experiment

Based on data gathered from three sources Twitter, Facebook, and personal conversations that cater to youth and young adults, the experimental results to validate the emotional score Children between the ages of 13 and 18 are impacted, while the young group consisted of those between the ages of 19 and 40. The studies are split into two phases, which are covered in more detail below, in order to assess the effectiveness of the suggested method.

1) Outcome

The empirical results of the suggested framework employing long words from the database for sensitivity analysis are shown in this subsection. The outcomes of the suggested and conventional sensitivity assessment methods for the paediatric group are displayed in Table 3. Long words are disregarded by the conventional technique, but the suggested algorithm incorporates them for a study of sensitivity. The efficiency of the suggested system is shown in black in Figures 6 through 9, which demonstrate a notable improvement over the conventional way. Compared to a traditional system, words achieve superior precision, recall, and F-measure rate (89.97%, 91.65%, and 90.8%, respectively) on the Facebook dataset. In comparison to the previous approach, the proposed system produces improved F-measures, namely 85.78% and 81.77 for the Twitter and Personal chat datasets, respectively. Additionally, Table 4 presents the outcome of the sentiment analysis method for the younger demographic. It is seen in Figures 6 through 9.

Figure 6: Comparison results of YouTube datasets for child group

Figure 7: Comparison results of YouTube datasets for young group.

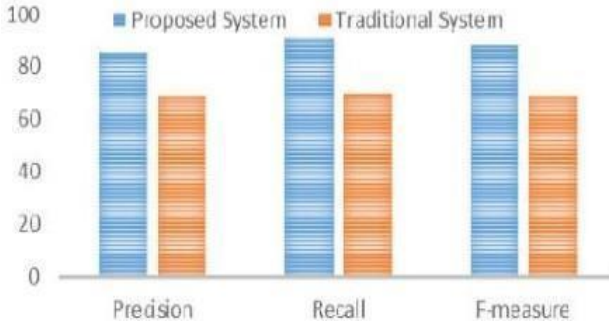

Figure 8: Comparison of chat datasets for child group

Figure 9: Comparison results of chat datasets for young group

***Corresponding Author:** Surya K M

Table.3: Comparison Results of Proposed and Traditional System of Children Group

Group	Dataset	Precision	Recall	F- measure
Proposed System	Facebook	89.97	91.65	90.8
	Twitter	84.56	87.04	85.78
	Personal chat	76.98	87.20	81.77
Traditional System(Ignore Lengthened words)	Facebook	69.78	65.67	67.47
	Twitter	64.34	66.88	65.05
	Personal chat	58.78	61.21	59.67

2) Effects of long words on perceptual evaluation
Because diverse scenarios were employed in the experiment, there is heterogeneity in the experimental outcomes of the suggested approach. Long words like "coolllll" and "happyyy" were also tallied in order to determine the final emotion score for the emotion evaluation. When capitalized phrases are added to the dataset, the suggested algorithm's average F-measure increases by 21.22% when compared to the old algorithm. Uppercase terms, in particular greatly enhance the novel scheme's performance.

The suggested system's effectiveness is assessed by comparing the outcomes to the gold standards. Gold standards are created with the assistance of four graduate students who rate comments, tweets and conversations according to emotion using a 7-point scale. Analysis of agreement among informants using distributed data examines the reliability of the degree of agreement and the feedback procedure. The hierarchy is analyzed using the kappa statistic, which shows that psychological intensity scores for children and adolescents are quite consistent with the gold standard.

The percentages of favourable and negative among the top 10 randomly selected papers are displayed in Figures 10 through 13. As demonstrated in Figures 10 and 12, there is not much of a difference between the outcomes produced by the suggested algorithm and the gold standard when compared to the conventional algorithm. Figure 13 illustrates that the findings for children are marginally smaller than those approaching the gold standard. The less frequent usage of lengthier words may be the cause of this discrepancy. However, the suggested structure shows how beneficial it is for young people's information sharing.

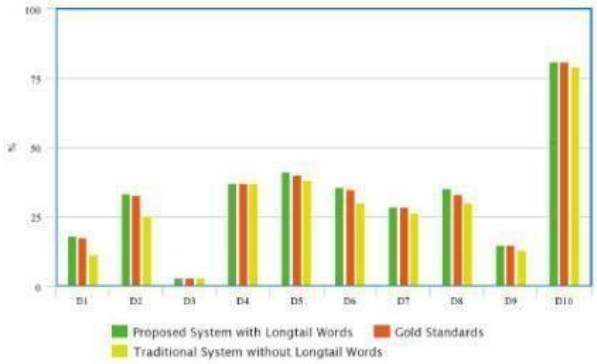

Figure 10: Effect Over Negative for Young

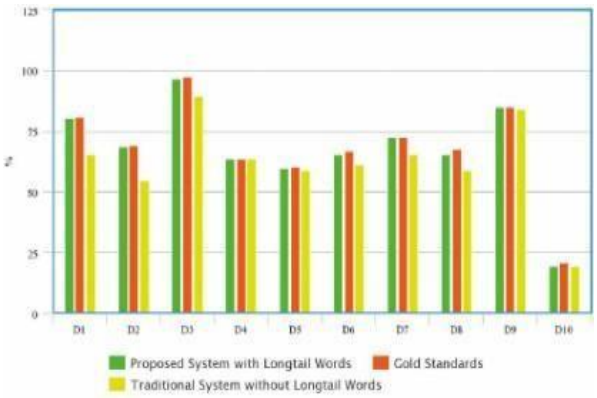

Figure 11: Effect Over Positive for Young

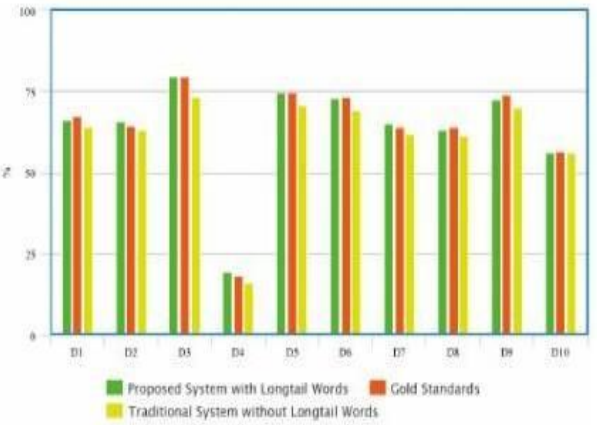

Figure 12: Effect Over Positive for Children.

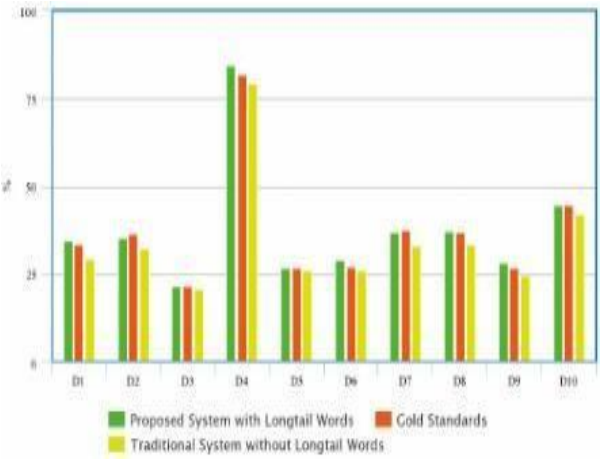

Figure 13: Effect over negative for young

***Corresponding Author:** Surya K M

Table 4: Performance measures of proposed and traditional methods of young group

Group	Dataset	Precision	Recall	F- measure
Proposed System	Facebook	96	94.51	95.25
	Twitter	91.45	89.43	90.43
	Personal chat	85.67	91.12	88.31
Traditional System (Ignore Lengthened words)	Facebook	76	75	75
	Twitter	71	69	70
	Personal chat	69	70	69

CONCLUSION AND FUTURE SCOPE

We introduced an emotion filtering method in this research that can be used for many tasks, such as emotion identification. It's critical to record the person's viewpoint in addition to their true emotions. Our research emphasizes how significant a role long words play in perceptual assessment. For more coverage, these terms need to be examined in more detail. As seen, elongation is frequently employed with subjective words to express meaning; it is not a random technique. Additionally, compared to older and younger age groups, the effect of time is positively associated with a rise in the percentage of young people who use the Internet.

Based on our research, we created an unsupervised length-based method to locate new words that affect emotions but aren't in dictionaries yet, as well as to ascertain the polarity of those words. Such tools are necessary in the ever-evolving net-speak micro blogging to understand modern tastes.

However, our proposed scheme has limitations. It takes correct spelling and is still producing results that are equivalent to the gold standard. Our goal is to improve and improve the efficiency of the system by developing new rules.

Only one facet of the elongation phenomena is investigated in this paper. We will look into additional word length characteristics in the future, like the connection between a word sample's perceptual intensity and length count.

Future research will look into delayed impacts in other languages, but for the sake of this study, we have concentrated on English cases. Finally, we intend to discuss elongation and cases where it is combined with other orthographic criteria to improve accuracy in a sentiment classification specific to Twitter. We will also explore the relationship between elongation events and other orthographic criteria, such as those related to perception and meaning, such as capitalization, perceptual cues, and punctuation.

REFERENCES

1. Patel, N. (2019, July 18). Longtail keywords: How-to, strategies, tips. Retrieved from https://neilpatel.com/blog/long-tail-keywords-seo/
2. Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253.
3. Salas-Zárate, M. D. P., Medina-Moreira, J., Álvarez-Sagubay, P. J., Lagos-Ortiz, K., Paredes-Valverde, M. A., & Valencia-García, R. (2016). Sentiment analysis and trend detection in Twitter. In *Proceedings of the International Conference on Technology and Innovation* (pp. 63–76). Cham, Switzerland: Springer.
4. Alashri, S., Kandala, S. S., Bajaj, V., Ravi, R., Smith, K. L., & Desouza, K. C. (2016, August). An analysis of sentiments on Facebook during the 2016 U.S. presidential election. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 795–802).
5. Montoyo, A., Martínez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, *53*(4), 675–679.
6. Probierz, B., Stefański, P., & Kozak, J. (2021). Rapid detection of fake news based on machine learning methods. *Procedia Computer Science*, *192*, 2893–2902.
7. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113.
8. Nguyen, B.-H., & Huynh, V.-N. (2022). Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach. *Journal of the Operational Research Society*, *73*(1), 102–121. https://doi.org/10.1080/01605682.2020.1784049
9. Kukkar, D., Mohana, R., Kumar, Y., Nayyar, A., Bilal, M., & Kwak, K.-S. (2020). Duplicate bug report detection and classification system based on deep learning technique. *IEEE Access*, *8*, 200749–200763.

***Corresponding Author:** Surya K M