International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

BOMEST: An Approach Big Data Analytics Framework for Sentiment Analysis of Unstructured Data Using SVM and NLP

Er. Aman Mishra [#1]

[#1] Assistant Professor, Dept. of CSE, Faculty of Engineering & Technology. KMC Language University

E- Mail: ***kmishra446@gmail.com***

ABSTRACT—

The exponential growth of data through sensors, ERP systems, and social networks, collectively described by the 3-Vs (volume, velocity, and variety), demands innovative data processing solutions. This research introduces the BOMEST Big Data Analytics Framework and Algorithm, designed for efficient sentiment analysis of unstructured data, particularly tweets from Twitter, using a lexicon-based approach. BOMEST enhances sentiment classification through part-of-speech tagging and real-time polarity scoring. The study evaluates BOMEST's application in cross-domain sentiment analysis and financial news sentiment classification, assessing the correlation between events and stock market returns. Experimental results demonstrate improved accuracy in sentiment classification and the potential for BOMEST in real-time analytics.

Keywords: - Sentiment Analysis, Big Data, NLP, SVM, Cross Domain Classification.

1. Introduction

Data generated from sensors, human-made applications, ERP systems, and social networks is commonly characterized by the "three V's": volume, velocity, and variety. This terminology reflects the vast scale and diversity of the information involved. To effectively harness big data, it is crucial to implement cost-efficient, innovative, and sophisticated data processing and analytics solutions. These solutions enhance insights, decision-making, customer satisfaction, and user experience.

With the emergence of the Social Web, individuals increasingly share their thoughts and ideas across various platforms, including blogs, tweets, and discussions on products or political issues. This trend necessitates the development of efficient algorithms to extract sentiments or opinions from the unstructured data generated by these sources. The process of summarizing an individual's views and sentiments regarding a specific entity is known as sentiment analysis, or opinion mining. [1] Natural Language Processing (NLP) plays a vital role in this context.

Twitter, being a significant source of big data, employs its REST API and OAuth credentials to retrieve tweets from hundreds of millions of users. This paper introduces the BOMEST Big Data Analytics Framework and Algorithm, which utilizes a lexicon-based approach to analyse the sentiment of each tweet and determine its polarity for real-time assessment.

The framework aims to identify both domain-specific and independent terms to enhance cross-domain data integration and minimize discrepancies. A method, in conjunction with the CDA algorithm, is proposed to improve sentiment classification. [2] BOMEST, an adaptation of Jain et al.'s algorithm, achieves a 78% accuracy rate within a single domain. It employs a bigram model to efficiently tag parts of speech (POS) and constructs a trained dictionary that catalogs various word combinations (e.g., noun + adjective, adverb + adjective). Sentiment polarity is computed by applying specific positive and negative weightings to the data, yielding a sentiment score ranging from 0.0 to 1.0.

Sentiment scores derived from news articles are critical for data analysis, particularly in the expanding field of news analytics. News reports about publicly traded companies often contain significant information that can influence stock prices. Therefore, analysing this data is essential for understanding its impact on stock market performance. Furthermore, sentiment analysis of financial news helps gauge reader sentiment, which can reflect the behavior of stock market investors. Investors' decisions are frequently shaped by their interpretations of news, which in turn affects stock returns, leading to discussions on market efficiency.

The structure of this chapter is systematic. Initially, two frameworks are proposed to address the issue, followed by a discussion of datasets and collection methods. Section 4.4 details the event-based sentiment analysis framework along with its results. [3] Section 4.5 applies Support Vector Machines (SVM) to predict news sentiment based on reader feedback, and the model is subsequently tested through trading strategies. The overall organization of the paper is as follows: Section 2 provides a literature review, while Sections 3 and 4 offer an in-depth explanation of the BOMEST Big Data Analytics Framework and Algorithm, focusing on tweets obtained from Twitter. Finally, Sections 5 and 6 conclude the research by validating the experimental results and presenting final insights.

2. Related Work

2.1 FRAMEWORK

Fig. 3.1 depicts our suggested BOMEST Big Data Analytic Framework for sentiment analysis of unstructured stream tweets connected to instructor. For each of the four levels of the structure, the following breakdown is provided: At this level of data storage, [4]Twitter Rest API OAuth token keys are used in Step 1 to collect tweets from search engines, the Web, and social networking sites like Facebook and Twitter. For this purpose, these tweets are sent to the message filter in Step 2 for data gathering based on that specific term. As illustrated in Fig. 4.1a, these tweets are translated into data in the form of text, audio, and video.

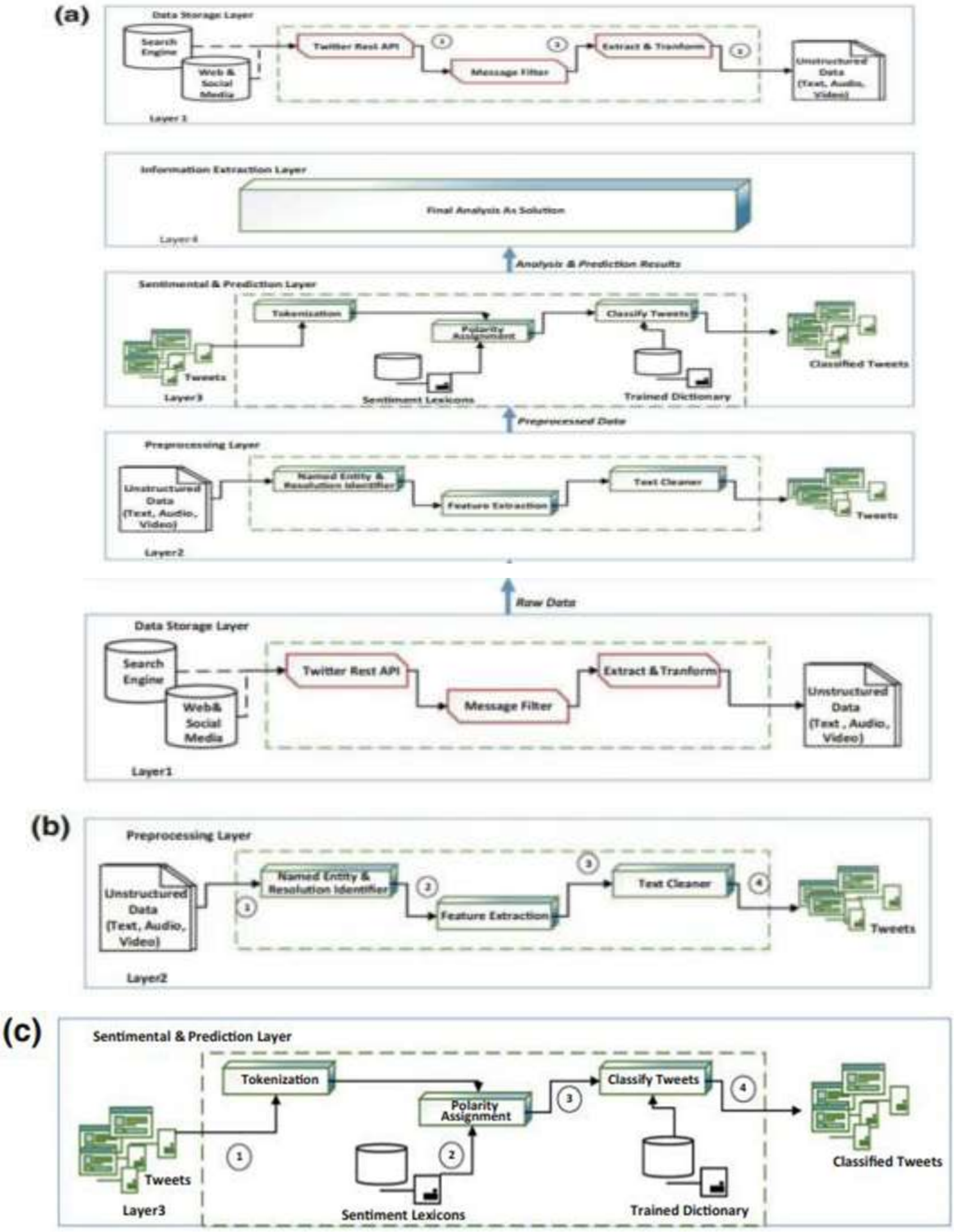


Fig. 2.1 BOMEST Big Data Analytic Framework a Data storage layer. b Preprocessing layer. c Sentimental and prediction layer

2.2 Information Extractions

Twitter Dataset: In the first phase, a dataset of 7,054 tweets tagged with the keyword "Teacher" was collected using the Twitter REST API. The Search API was used to gather tweets in real-time that contained the term "teacher" or phrases related to teachers. [5]The implementation of this experiment was done using the .NET framework. Figure 2.2 illustrates the code used to retrieve tweets from Twitter.

Data Cleaning: This stage focuses on eliminating inconsistencies and irrelevant content from the tweets to improve the quality of the dataset. Figure 2.3a highlights that some tweets consist only of hyperlinks, which do not provide meaningful information. To enhance the dataset, these tweets are removed before processing. A novel algorithm was developed to automatically delete all tweets that contain any kind of link (such as videos or images) and remove punctuation marks (such as @, :), commas, etc.). After filtering out unnecessary content, blank spaces, and URLs, the cleaned tweets are ready for further analysis.

2.3 Dataset (data collection/ Refining data) -Twitter Data set

The dataset is organized into three distinct collections. The first collection contains a list of companies along with their ticker symbols, index associations, and countries of origin. This data was gathered from online sources and includes information on over 4,500 stocks from the US, UK, and Germany, across various indices.The second collection stores daily records of stock market prices for each company, including statistics such as the opening price, highest and lowest prices, closing price, and trading volume.The third collection consists of more than 120,000 news articles. Each news entry includes details such as the publication date, publisher, HTML page, and the date the article was crawled, providing a basis for illustrating how raw data is processed.

2.4 Data Collection

In this section, news stories spanning from January 6, 2011, to September 9, 2011, are examined. News articles were collected daily using Yahoo's Query Language (YQL) from multiple sources and stored for later analysis.

It’s important to note that this thesis focuses on formal news, which includes topics presented by credible news outlets, whether broadcast on television, radio, or published in print or online. [6]The news content is obtained through services and APIs provided by Yahoo!. Initially, the study concentrated on FTSE100 companies based in the UK, with global index data being gathered in the first phase of the investigation.

The raw data is processed by first removing HTML tags. Afterward, relevant news items, along with their publication dates and titles, are extracted and stored in a news database. The data is then refined for analysis using two different methods: Natural Language Processing (NLP) and Open Calais are employed to extract events and associated information for event-based sentiment analysis

2.5 EVENT BASED SENTIMENT ANALYSIS

This section explores whether there is a correlation between news stories and stock market movements. The first step involves retrieving relevant events from the news. Semantic information can be extracted from text using Open Calais, which provides a Java API for access. When given a text to analyse, Open Calais returns a range of information, including Fact Index and Definitions, Entity Relevance Scores, and Entity Tags.

Open Calais identifies several occurrences that could influence market outcomes. These events are customized based on the data provided for the investigation. [7] However, it is not feasible to gather a sufficient sample of news for events with low occurrence frequency, and therefore, such events will not be further analysed. The Open Calais API returns data in four formats: microformats, basic format, JSON, and N3. Since these formats vary significantly, JSON has been chosen as the preferred format for this thesis.

The Java Calais Parser class is used to extract relevant data. For instance, a JSON parser library was utilized to convert the JSON response into a list of paragraphs, each containing useful information. An example of event extraction is shown in Listing 2.1.

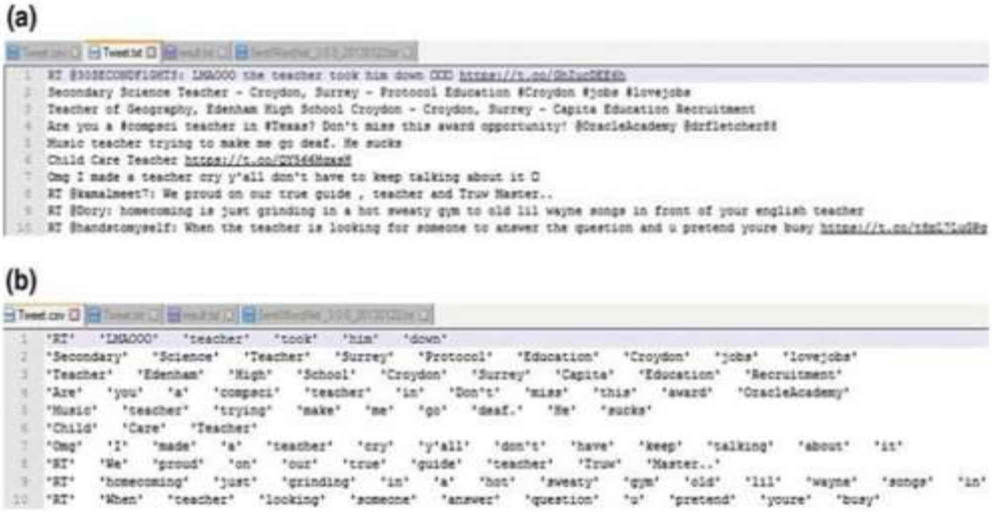
Listing 2.1: Example of Information Extracted from JSON

Company Customer	Employment Relation	Product Release
Equity Financing	Secondary Issuance	Product
Programming Language	Company Expansion	Family Relation
Trial	Province Or State	Company Force Majeure
FDA Phase	Voting Result	Published Medium

3. Proposed Work

3.1 Proposed Work Methodology:

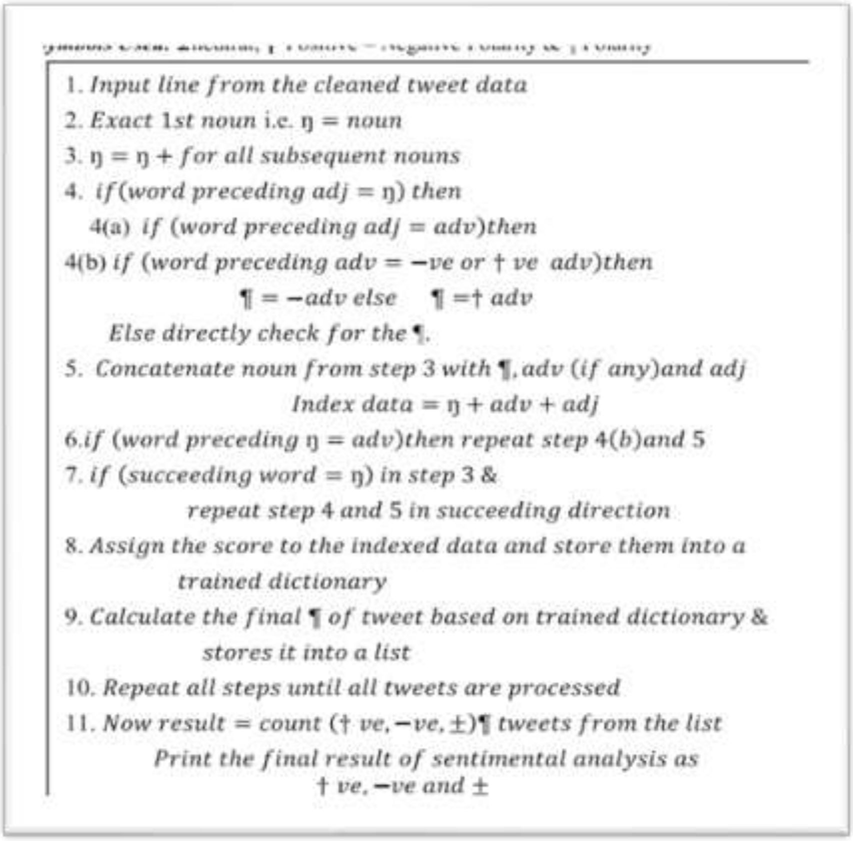
As a consequence, we developed a new algorithm, the BOMEST, which uses a clean dataset as input and provides greater POS tagging efficiency, assigns polarity scores and improves the accuracy of Twitter tweet analysis by evaluating each tweet and extracting information in real time, as shown in Fig 3.1.



```
RT @30SECONDFIGHTS: LMAOOO the teacher took him down ☐☐☐ https://t.co/GhIucDEE6h
Secondary Science Teacher - Croydon, Surrey - Protocol Education #Croydon #jobs #lovejobs
Teacher of Geography, Edenham High School Croydon - Croydon, Surrey - Capita Education Recruitment
Are you a #compeci teacher in #Texas? Don't miss this award opportunity! @OracleAcademy @drfletcher88
Music teacher trying to make me go deaf. He sucks
Child Care Teacher https://t.co/GY544Mqxs8
Omg I made a teacher cry y'all don't have to keep talking about it ☐
RT @kamalmeet7: We proud on our true guide , teacher and Truw Master..
RT @Dory: homecoming is just grinding in a hot sweaty gym to old lil wayne songs in front of your english teacher
RT @handstomyself: When the teacher is looking for someone to answer the question and u pretend youre busy https://t.co/tKnL7LuGRg
```

```
'RT'    'LMAOOO'    'teacher'    'took'    'him'    'down'
'Secondary'    'Science'    'Teacher'    'Surrey'    'Protocol'    'Education'    'Croydon'    'jobs'    'lovejobs'
'Teacher'    'Edenham'    'High'    'School'    'Croydon'    'Surrey'    'Capita'    'Education'    'Recruitment'
'Are'    'you'    'a'    'compeci'    'teacher'    'in'    'Don't'    'miss'    'this'    'award'    'OracleAcademy'
'Music'    'teacher'    'trying'    'make'    'me'    'go'    'deaf.'    'He'    'sucks'
'Child'    'Care'    'Teacher'
'Omg'    'I'    'made'    'a'    'teacher'    'cry'    'y'all'    'don't'    'have'    'keep'    'talking'    'about'    'it'
'RT'    'We'    'proud'    'on'    'our'    'true'    'guide'    'teacher'    'Truw'    'Master..'
'RT'    'homecoming'    'just'    'grinding'    'in'    'a'    'hot'    'sweaty'    'gym'    'old'    'lil'    'wayne'    'songs'    'in'
'RT'    'When'    'teacher'    'looking'    'someone'    'answer'    'question'    'u'    'pretend'    'youre'    'busy'
```

Fig. 3.1 a Twitter data before cleaning. b Twitter data after cleaning



```
1. Input line from the cleaned tweet data
2. Exact 1st noun i.e. $ŋ = noun$
3. $ŋ = ŋ +$ for all subsequent nouns
4. if(word preceding adj = ŋ) then
   4(a) if (word preceding adj = adv)then
4(b) if (word preceding adv = −ve or † ve adv)then
                 ¶ = −adv else    ¶ =† adv
     Else directly check for the ¶.
5. Concatenate noun from step 3 with ¶, adv (if any)and adj
                 Index data = $ŋ + adv + adj$
6.if (word preceding ŋ = adv)then repeat step 4(b)and 5
7. if (succeeding word = ŋ) in step 3 &
          repeat step 4 and 5 in succeeding direction
8. Assign the score to the indexed data and store them into a
          trained dictionary
9. Calculate the final ¶ of tweet based on trained dictionary &
          stores it into a list
10. Repeat all steps until all tweets are processed
11. Now result = count († ve, −ve, ±)¶ tweets from the list
          Print the final result of sentimental analysis as
                 † ve, −ve and ±
```

The CDA is a created algorithm that employs Amazon review data (reviews of baby, beauty, health, and electronic items) as an input dataset and is more efficient than BOMEST. As a classifier must be trained for each new domain, analyzing several domains takes a long time and is expensive. As a result, a new method is required, one that can be used across domains. A cross-domain classifier is used to improve the current approach by merging two separate source domains and predicting the outcomes of the target domain using the steps outlined in the algorithm.

3.2 READER-BASED SENTIMENT ANALYSIS

This section proposes a novel approach for classifying news emotion. Afterwards, it employs the methodology proposed in Chapter 3 to examine the model's capacity to forecast the news. In order to prepare for the categorization procedure, the news is first categorized as either good or negative.

3.3 REFINING THE DATA

Data pre-processing is carried out using the methodology proposed in Chapter 3. According to distinct χ2 values, the F (f) matrix is built using the feature presence (FP) weighting approach. For each characteristic, Fig. 3.3 depicts the relationship between the χ2 statistic and the frequency F (t) across all documents. Features in the top right corner of these panels are obviously the most desired, since one is looking for a trait that is usually proven to be a good predictor of the outcomes The left-hand panel, on the other hand, illustrates that such a mix is very unusual. [8] Low explanatory power (χ2) and low frequencies (F (t)) dominate the bottom left-hand side. An impressive number of characteristics may be used to explain the observed behaviour. In the right-hand panel, they are shown. To the untrained eye they seem like several terms that a human reader may associate with feelings of resentment like "victim," "corrupt," and "violate."

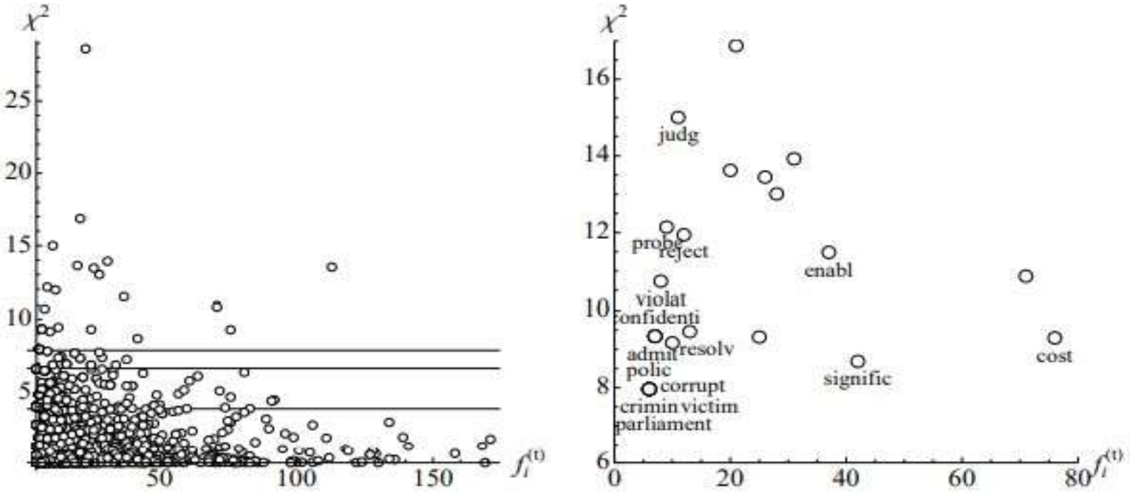


Fig 3.2. Scattered plot

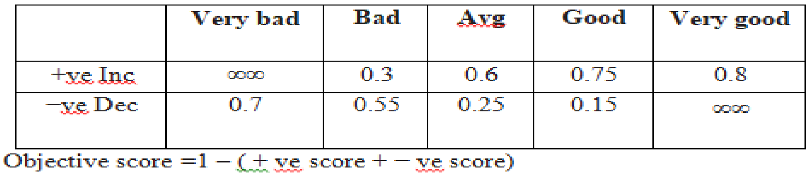
Scattered plot in Figure 4.7 shows correlation between frequency of feature fi and 2 test statistic significance in classification method. Starting with the lowest p-value, p = 0.05(χ2 = 3.84), p = 0.01(χ2 = 6.63), and p = 0.005(χ2 = 7.88), the three vertical lines on the left represent the three distinct p-values.

Upper-right traits are the most sought after. They may be seen in a zoomed-in perspective on the right-hand side graph.

4. RESULT AND PERFORMANCE ANALYSIS

Two algorithms are used in our sentiment analysis studies here. First, the tweets are tagged with POS using the Bag-of-Words and Sent WordNet vocabulary. To get an objective assessment, the dictionary gives each synset a score of either +1 or -0, depending on how positive or negative the synset are:

Table 4.1: Score of new synset based on (+, −) polarity



	Very bad	Bad	Avg	Good	Very good
+ve Inc	∞∞	0.3	0.6	0.75	0.8
−ve Dec	0.7	0.55	0.25	0.15	∞∞

Objective score =1 − (+ ve score + − ve score)

Bag-of-Words is used to tag all the POS in the second experiment. The BOMEST algorithm is based on a bigram model, and it tags all the POS in an efficient way, as explained in Section 4. It also uses a trained dictionary that stores all the combinations of words of nouns + adj and adv + adj and other words. the number of points for each new synset is added together. Indexed data shows that +ve polarity increases the value by 0.45, while -ve polarity decreases it 0.35 for the indexed data. [10]The sum of all of these is in the range of 0.0–1.0, as shown in Table 5.1.

In Fig. 5, you can see how the two algorithms work together. This is shown by the fact that the more words in the dictionary, the better the sentiments are. If a large dictionary with effective scoring schemes is used, it may improve accuracy. This is because the dictionary is dense and the scoring schemes are effective. So, our trained dictionary has an accuracy rate of more than 78% when it is run through a BOMEST algorithm.

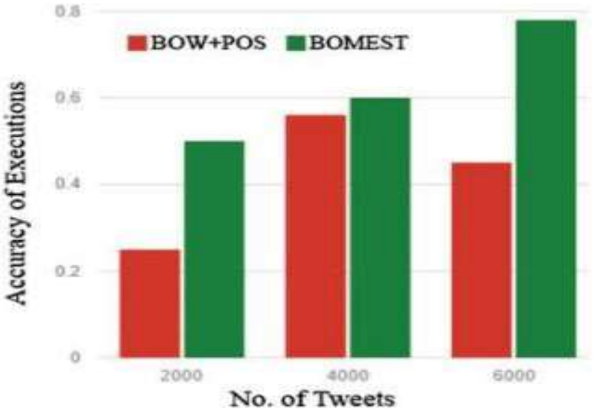


Fig. 4.1: Comparison of BOW + POS, BOMEST

To begin, the Naive Bayes approach is used to see whether the daily stock market returns are affected by the presence of an event. For each event, the likelihood of a positive return is determined based on the chance of that event occurring. Starting and closing prices of time t is used to compute returns. Let E ={e1, e2, . . ., en} denote a set of the possible events that might occur, and et is a binary variable denotes the occurrence of event n in time t

$$e_{n,t} = \begin{cases} 1, & \text{the event } n \text{ occurred} \\ 0, & \text{else.} \end{cases}$$

Then P(r+|en) denotes the probability of positive returns, given an event n has occurred:

$$\mathbb{P}(r^{+}|e_n) = \frac{\sum_{t=1}^{T} e_{n,t}\, r_t^{+}}{\sum_{t=1}^{T} e_{n,t}},$$

where r+ ∈ {1, 0} depends on whether the return is positive or not. The values of
t
P(r+|en) for all events are illustrated in Fig. 4.5.

Fig. 4.2 demonstrates that a number of occurrences are closely associated with high profits. Event "al- liance" attracts attention because of the high chance of a positive return given the occurrence of this event, which is 82%. Accordingly, low percentages are also interesting since they indicate a greater likelihood of negative returns.

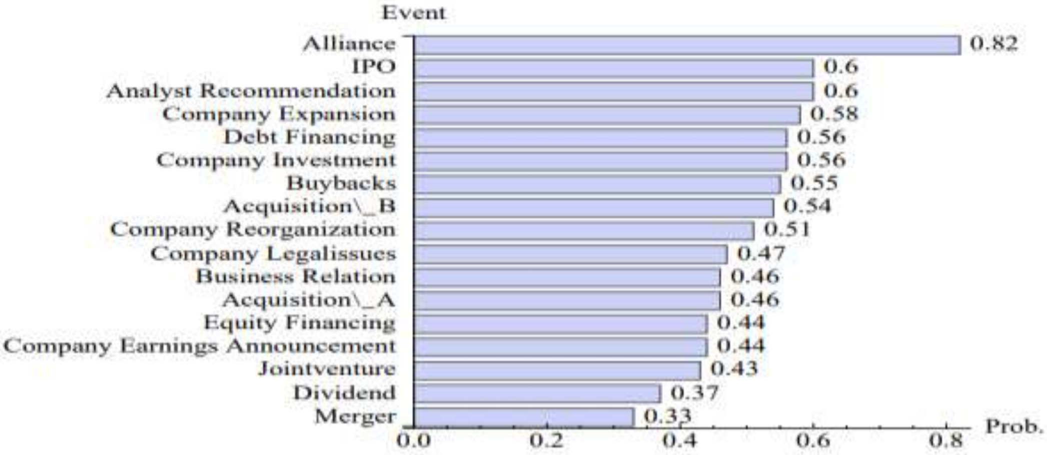


Figure 4.2: Conditional Probabilities of Rising Returns

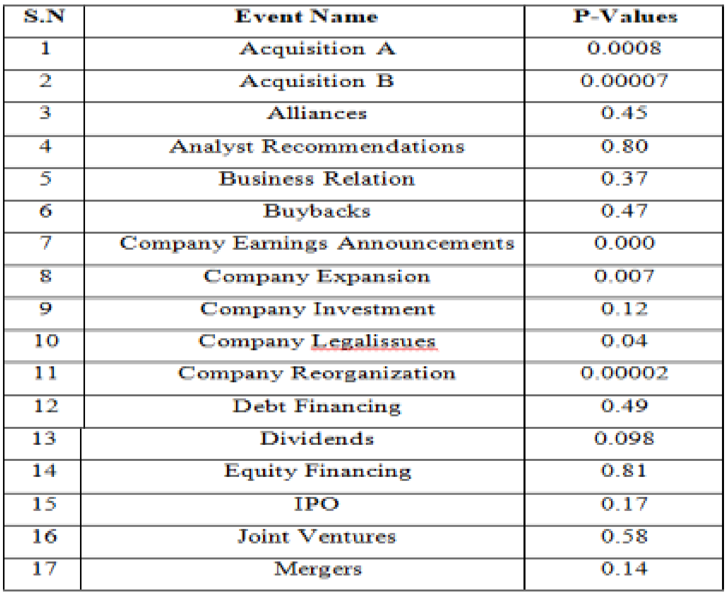
when the events occur, because P(r−|en) = 1 − P(r+|en).

Consider for instance the event “merger”, P(r+|emerger) = 0.33 implies that

P(r−|emerger) = 0.67. That indicates that given that the event is “merger”, the prob- ability that the returns is negative is 67%.

To gain insight into which events might influence stock returns, we can start by using the Naive Bayes approach. However, for a more accurate assessment of the relationship between returns and events, a more detailed analysis is needed. Some events show no significant impact, such as "company reorganization." For example, P (r+|ecomp.reorg.) = 0.51 and P (r−|ecomp.reorg.) = 0.49 indicate that this event does not have a notable effect on stock returns. In contrast, the event "alliance" shows a stronger correlation, with P (r+|ealliance) = 0.82. A statistical test is needed to determine whether this deviation from a 50/50 probability is statistically significant. If appropriate, a t-test will be conducted to confirm this. If the data does not meet the assumptions of parametric testing, the non-parametric Wilcoxon Rank test will be used instead. The two tests are applied separately, with parametric tests preferred when possible and non-parametric tests used only when necessary.

Table 4.2: Results of Jarque-Bera Test



S.N	Event Name	P-Values
1	Acquisition A	0.0008
2	Acquisition B	0.00007
3	Alliances	0.45
4	Analyst Recommendations	0.80
5	Business Relation	0.37
6	Buybacks	0.47
7	Company Earnings Announcements	0.000
8	Company Expansion	0.007
9	Company Investment	0.12
10	Company Legalissues	0.04
11	Company Reorganization	0.00002
12	Debt Financing	0.49
13	Dividends	0.098
14	Equity Financing	0.81
15	IPO	0.17
16	Joint Ventures	0.58
17	Mergers	0.14

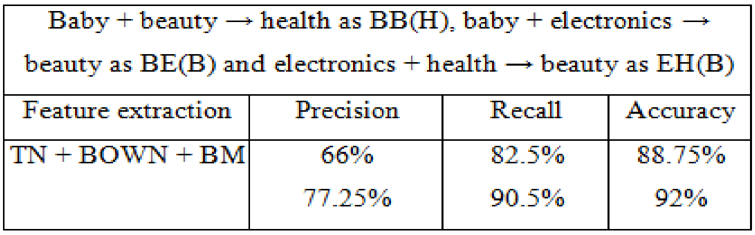
The null hypothesis for the t-test is that the sample mean equals zero, and the test can only be conducted if the sample follows a normal distribution. To verify normality, the Jarque-Bera (JB) test will be used. The JB test examines whether the sample data is normally distributed, and the results for return samples corresponding to different events are summarized in Table 4.2. [10] For events such as analyst recommendations, business partnerships, stock buybacks, corporate investments (in both debt and equity financing), dividends, and initial public offerings (IPOs), the null hypothesis cannot be rejected as the test results show p-values greater than 0.05. Consequently, the t-test will be applied for these events. For those events where the p-values are less than 0.05, indicating that the null hypothesis is rejected, the non-parametric Wilcoxon signed-rank test will be applied.

In this experiment, the source domain was composed of a random combination of two domains, while the remaining dataset served as the target domain. For example, terms such as "Baby + Beauty → Health" (BB(H)), "Baby + Electronics → Beauty" (BE(B)), and "Electronics + Health → Beauty" (EH(B)) were used when calculating precision, recall, and accuracy metrics. To identify domain-independent terms from the source domain, an estimate metric was employed, functioning as a cross-domain classifier. The classifier's performance was evaluated in terms of Precision (P), Recall (R), and Accuracy. (A) are determined based on the results of the following outcomes.

It is defined as: $\frac{tp}{fp+tp}$. Recall ® determines the completeness and integrity of the classifier. It is defined as: $\frac{tp}{f\eta+tp}$. Accuracy (A) determines how often a sentiment is close to positive or negative sentiments. It is defined as: $\frac{tp+t\eta}{PS+NS}$.

Table 4.3 clearly shows that the CDA's recall and accuracy metrics are higher than BOMEST's when compared. These results were achieved using TN, BOWN, and BM (BOMEST): Tokenization 66%, 82.55% recall and 88.75% accuracy. As demonstrated in Fig. 4.3., BM, WSR, and CBM (Cross BOMEST) generate 77.25 percent precision, 90.5 percent recall, and a maximum accuracy of 92 percent, which is greater than that of BOMEST (i.e., 81%).

Table 4.3: Precision, recall, and accuracy results



Baby + beauty → health as BB(H), baby + electronics → beauty as BE(B) and electronics + health → beauty as EH(B)			
Feature extraction	Precision	Recall	Accuracy
TN + BOWN + BM	66%	82.5%	88.75%
	77.25%	90.5%	92%

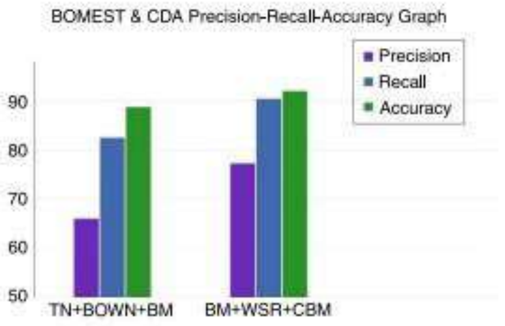


Fig. 4.3: Precision–recall–accuracy graph

4. CONCLUSION

To derive valuable insights from large volumes of raw data, cost-effective, innovative, and creative information processing and analytical solutions are essential. In this study, the BOMEST technique was developed using sentiment analysis to extract relevant information. Specifically, BOMEST was applied to analyze tweets. The preprocessing phase began with data cleaning, followed by Part-of-Speech (POS) tagging, which assigned polarity to the processed data. Tokenization was then carried out, assigning positive, negative, or neutral values to each token. These token values were ultimately used to generate the final output. Implementation results show that the proposed BOMEST lexicon method achieved 78% accuracy, significantly outperforming the existing lexicon technique, which had an accuracy of 46%.

This chapter also explores financial news sentiment using both event-based and reader-based techniques. Event-based sentiment analysis gauges public sentiment regarding a news story based on reported events. To achieve this, the relationship between a company's stock returns and reported events was examined using both parametric and non-parametric statistical methods. Using a simple Naive Bayes approach, a strong correlation was found between stock returns and an "alliance" event, as illustrated in the graph. Additional tests were conducted to determine if specific events had a statistically significant impact on stock prices. These tests confirmed that the Naive Bayes model was correct: the "alliance" event significantly affects stock returns. To further validate this, a regression model was applied. However, since most events did not correlate with stock market returns, the event-based model could not be conclusively proven. This conclusion may vary depending on the data. Therefore, the potential effectiveness of the event-based sentiment model cannot be dismissed entirely. A larger dataset with more frequent occurrences of events might reveal stronger correlations with stock market gains. Moreover, a non-linear model could potentially describe the evolution of this relationship more effectively. Support Vector Machines (SVM) were used to classify sentiment in financial news and assess the accuracy of this approach.

Additionally, a Cross-Domain Sentiment Analysis (CDA) method was proposed, which involved creating the Lexical Boms Dictionary. After removing irrelevant data and applying stemming, this dictionary was employed to enhance positive reviews. These reviews were then converted into tokens using the Bag of Words (BOW) and BOMEST methods. The positive and negative reviews were indexed, and synonyms were replaced with corresponding terms to amplify the polarity in search engines. To evaluate performance across different domains, two separate source domains were trained to extract reviews, and CDA was used to identify terms that could bridge the gap between texts from diverse domains. Various techniques such as BMs, WSRs, CBMs, target extraction, and cross-domain classifiers were employed. The CDA method achieved 92% accuracy when applied across domains, while BOMEST showed an improvement of 16% in accuracy and 7% in recall within a single domain. Compared to existing methods, CDA improved both precision and accuracy by 5%.

Although this study did not explore cross-domain implementations, future research aims to apply the proposed method to a wider range of contexts. Future work will also focus on optimizing the algorithm's time and space complexity when compared to current techniques.

5. REFERENCES

[1] Jain V., et al., "BOMEST a Vital Approach to Extract the Propitious Information from the Big Data," LNNS Springer's, 2016. [

[2] Pan S., et al., “Cross Domain Sentiment Classification via Spectral Features,” in WWW 2010, Raleigh, North Carolina, USA. ACM, pp. 751–760, 2010.

[3] Hu M. and Liu B., “Mining and summarizing customer reviews,” in KDD, pp. 168–177, 2004.

[4] Pang B. and Lee L., “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval, pp. 1–135, 2008.

[5] Lu Y., et al., “Rated aspect summarization of short comments,” in WWW, pp. 131–140, 2009.

[6] Blitzer J., et al., “Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification,” in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 440–447, 2007.

[7] Pan S., et al., “Cross Domain Sentiment Classification via Spectral Features,” in WWW 2010, Raleigh, North Carolina, USA. ACM, pp. 751–760, 2010.

[8] Fang H., “A re-examination of query expansion using lexical resources,” in ACL, pp. 139–147, 2008.

[9] Bollegala D., et al., “Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus,” IEEE transactions on knowledge and data engineering, 2013.

[10] Chen B., et al., “Extracting discriminative concepts for domain adaptation in text mining,” in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM, pp. 179–188, 2009.