International Journal of Modern Agriculture, Volume 9, No.3, 2020
ISSN: 2305-7246

A Study on the Sentiments of the Indian Public towards the Budget Using a Machine Learning Algorithm

Aarti Mehta Sharma [1], Nishigandha Jangnure [2]
[1]Symbiosis Institute of Business Management, Bengaluru, Symbiosis International (Deemed University), Pune, India
[2]Cummins India Ltd, Pune, India
Email: [1]aartimehtasharma@sibm.edu.in,[2]nishigandhajangnure@gmail.com

Abstract

Purpose: A lot of hope and aspiration is hinged to economic events like the presentation of the Union Budget by the government as it is expected to bolster economic growth and foster social justice and equity. The highlights of the Indian Budget 2019 were that India is expected to become a US$ 3 trillion economy by the end of this financial year 2019-2020. The paper proposes to ascertain if the public was happier after the budget, lending credence to the second term of the present government

Design/methodology/approach: Various machine learning techniques using Sentiment Analysis have become popular to gauge twitterati’s reactions to major events in the country. This paper proposes a hybrid approach to sentiment analysis with probabilistic topic modelling (using Latent Dirichlet Allocation (LDA)) to examine subjects pertaining to the budget that were most discussed by twitterati during and after the budget. We will further assign sentiments to these topics by using Naïve Bayes algorithm for classification into positive or negative sentiment. The programming software R has been used to extract and analyse tweets. A total of twenty two thousand tweets have been extracted for this study.

Findings: It was found that people were not pleased with the union budget.

Originality: This paper explores topic modelling using LDA and shows that tweets can be used to gauge the public’s mood before and after the announcement of an event.
Keywords: unionbudget, sentiment analysis, topic modeling, LDA, latent dirichlet allocation, twitter, R software, social media analytics, policy changes, tweets, machine learning

Introduction

The Union budget is a document published by the Government of India and it states the various sources of funds for the Indian government along with a tentative outline of how it plans to use these funds to achieve development goals. As the budget determines the source and allocation of public funds, it has a huge economic, social and political impact. With respect to economic policymaking, the union budget is the most-watched event in India. Policies regarding taxation, subsidies, expenditures and investments are important from the macroeconomic point of view and are outlined in the budget. The government may use the budget speech as a medium to give out information regarding the new policy initiatives that it plans to undertake. For the financial year 2019-2020, taxes formed the primary revenue source while significant spending was for the states' shares of taxes and duties and on interest payments against borrowings. Areas of government spending have a direct bearing on the lives of citizens. If a significant part of the revenue is spent on interest payments, it leaves little money for critical public expenditures like welfare schemes, defence, research, and development (Source: https://www.indiabudget.gov.in/budget2019-20/index.php).

In today's age of web 2.0, reactions of people on social networking services have become a barometer of public sentiment. People use websites like Facebook, Twitter, and Instagram extensively to express themselves. Every day the authors of messages on these platforms post about their life, discuss the latest issues, and give their opinions about a variety of topics. As a result, social media platforms have become a rich source of data, which can be used to know about people's sentiments and opinions. Twitter is one such extremely popular social media website. Since its inception in 2006, the number of twitter users has proliferated. Today there are 321 million active users as of February 2019. On Twitter, people make posts in the form of short texts called tweets. These tweets are limited to a maximum of 280 characters.

Though it is possible to post on Twitter in the form of images and videos, text posts are dominant. The importance of text analytics has thus increased multifold. Sentiment analysis enables the segregation of text into positive, negative, and neutral opinions (Pang, 2008). Whenever people use a word or a phrase extensively in their tweets, it appears as a twitter trend. Trends are the topics being discussed by people in a particular geographical area at a particular instant (https://help.twitter.com/en/using-twitter/twitter-trending-faqs). However, it is not possible to track these trending topics over an extended span of time. In this paper, we propose a model that uses topic modeling, Naive Bayes classification, and dictionary-based sentiment analysis to look at the changing sentiment of people towards a particular issue.

Macmillian English Dictionary defines 'sentiment' as 'a belief or an attitude towards something.' In the context of machine learning and text analytics, sentiment analysis is "The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic or product is positive, negative, or neutral." (https://www.lexico.com/en/definition/sentiment_analysis, Oxford). Table Ⅰ gives examples of each of these types of sentiment:

Table 1: Examples of sentiments

Glad to see EV push from the budget!	Positive
Budget 2019 allocates Rs 70,000 crore to PSU banks. Will the banking sector funds gain?	Neutral
The farmers in the country are suffering. I want to draw the government's attention towards it. No concrete steps were taken in the Union Budget to provide relief to the farmers	Negative

Sentiment analysis can be used to annotate the text as one of the following - positive, negative, and neutral. Three levels of sentiment analysis are presently based on the text type on which it is performed- document level SA, sentence-level SA, and aspect level SA. The document-level sentiment analysis is useful when analyzing long text like books and blogs. There is not much difference in the document, and sentence level SA as sentences can be treated as small documents (Liu, 2012). Sentence level SA is used for analyzing the sentiment in comments, tweets, etc. The aspect level SA makes use of particular product features that may be mentioned in the text. This kind of sentiment analysis is useful, particularly when we are looking at an opinion about a particular product feature.

The bag of words approach is often the basis of sentiment analysis. The problem with this kind of approach is that the document term matrix is created using the entire vocabulary of the corpus text. Such a matrix can become huge in size and may lead to sparsity problems. Our paper tries to resolve this problem by using topic modeling. The topic words obtained from LDA are used as the terms in the document term matrix with Naive Bayes algorithm. This is expected to make the algorithm computationally lighter while preserving or improving its classification ability.

Related work

Most of the scholarly work has been done in analyzing the public support for government spending in various areas (Ladd,1979 ; Maslauskaite, 2011). Shifts in the political orientation of the citizens and the change over a time period have also been attempted through analysis of government spending (Davis, 1980). Besides, research has been done on the benefits of public participation in government budgeting, also called participatory budgeting. Through participatory budgeting, people's opinions about the spending of public funds are taken into account. This has shown increased public engagement in governance as well as increased social inclusion. Table Ⅱ outlines some of the research done in government policy spending and public reaction to it. (Jacoby, 2000; Maslauskaite, 2011; Ram, 2008)

Table 2. Research on government spending

Authors	Paper Title	Year	Gist
Andre´s Rodrı´guez-Posea, and Kristina Maslauskaite	Can policy make us happier? Individual characteristics, socio-economic factors and life satisfaction in Central and Eastern Europe	2011	A study of the citizens of Central and East European countries shows that institutional factors like corruption, Government spending and a decentralized form of government contribute to overall wellbeing and happiness of people.
Christopher Ellis & Christopher	Social Policy and Public Opinion: How the Ideological Direction of Spending Influences Public Mood	2011	The electorate is not as reactive to the total spending by government as it is to the allocation of funds to direct and

Faricy			indirect spending. Changes in pattern of spending are seen as ideological shifts in policy making by the people.
Rati Ram	Government spending and happiness of the population: additional evidence from large cross-country samples	2008	There is no negative correlation between the happiness of people and size of government spending. This is in contradiction to some earlier scholarly work.
William G. Jacoby	Issue Framing and Public Opinion on Government Spending	2000	The way everyday issues are presented to the public, the choice of words etc. have a great effect on the support gained for various government spending programs. The paper explores effect of framing on political support received from public.
Ladd et al	The polls : Taxing and Spending	1979	Views of Americans on government spending and taxation. People are dissatisfied on the amount of tax and the (unfair) inequality in proportion of tax their income group has to pay. 74% of Americans were dissatisfied with the taxation which led to tax revolt in 1978. There is opposition to taxation but want of spending by government for services (source of these funds is taxes). Citizens want the government to be prudent about its spending by reducing waste generated by bureaucracy.

Of late, social media is being used extensively by people to express themselves. Thus, one of the reliable sources of public opinion is social media platform, Twitter. Extensive work on sentiment analysis has been done previously in the fields of movie reviews (Maas, 2011) (Pang, 2002), product review (Gulla, 2010), and twitter and other micro-blogs (Go,2009). The use of opinion mining is on the rise in the field of government intelligence and politics, as well. Sometimes, it is used to know about political opinions of the citizens (Singh ,2010) while sometimes text mining may be used to know the stance of a political figure or government towards an issue (Bansal, 2008 ; Laver, 2003 ; Ficcadenti, 2019). This helps the voters make a better choice by enhancing the information available to them. Sentiment analysis and opinion mining are finding increasing use in e-governance (Bansal, 2018 ; Kwon, 2006). Due to the overwhelming amount of data available through social media, governments are contemplating the use of such information to crowd-source ideas and opinions of people. An example of this in the Indian context was seen when the Indian Finance Ministry invited ideas from the general public for drafting the union budget 2019. Sentiment analysis has been used multiple times as an attempt to predict the outcome of political events like elections.

Sentiment analysis systems can be grouped into two groups: Machine Learning approach and Lexicon based approach. Various algorithms like Naive Bayes, Support Vector Machine (SVM), Maximum Entropy, k Nearest Neighbor (KNN), Artificial Neural Networks (ANN), etc. are used in supervised learning methods. In the most cited paper on sentiment analysis (Pang, 2002), the authors provide a comparative analysis of Naive Bayes, SVM, and Maximum Entropy methods. They use various features like unigrams, bi-grams, Parts of Speech, combinations of these, etc. The paper concludes that unigrams perform better than bi-grams, and the presence of frequent terms gives better results than the frequency of terms. Amongst the various classification algorithms, the Naive Bayes algorithm is shown to give the best results when the sample size is small (Hartmann, 2019). In our model, we thus use the Naive Bayes algorithm for classifying the tweets into various topics identified through topic modeling.

In Lexicon based methods, the tweet is classified as positive, negative, or neutral based on a predefined bag of words as per their polarity. The most popular lexicons in use are SentiWordNet, SentiWords, AFINN, etc. These word databases are dictionaries of words with their polarities defined. This approach is fast and requires no labeled data, but it has its own limitations. The word polarities defined in these lexicons are not domain-specific.

E.g., This movie is funny - Positive sentiment

This sandwich has a funny taste - Possibly negative connotation

Using Lexicon based method for polarity detection may lead to incorrect classification as a single word could denote different emotions in different contexts. The limitation of this approach is the unavailability of domain-specific databases.

Sentence level polarity detection is essential, but policymakers might not be able to know which topics are of utmost concern to the people. They might not be able to decipher the sentiment people associate with a particular issue. Thus, detecting the polarity of a topic can provide valuable insights into public opinion.

Latent Dirichlet allocation (Blei, 2003) is a machine learning method most commonly used for explorative topic identification (Zhang, 2017 ; Puranam, 2017). This method discovers a predetermined number of topics from the corpus.

Our paper provides unique work in the following ways:

- Using topic words from topic modeling for dimensionality reduction of the Naive Bayes algorithm. This new method of feature extraction is expected to improve the performance of the classification model. Such a method has not been explored previously to the best of our knowledge.

- Sentiment analysis of topics as a whole is another unique feature our paper tries to explore.

Research Objectives

The main objectives of this research paper are:

- To identify the topics trending during the week the Union Budget was announced
- To monitor the change in the public sentiment towards these topics before and after the budget declaration
- To introduce a novel methodology to perform automated text classification with less resources while maintaining the quality of output

Methodology

We downloaded 5000 tweets using the Twitter API.Hashtags like #UnionBudget, #Budget2019, #BudgetForNewIndia etc. were used to extract the tweets. To take into account the effect of sentiment change before and after the declaration of budget, the tweets were collected over the week when the budget was announced. Thus, tweets from 3 July 2019 to 12 July 2019 were used for our study. While extracting the tweets, care was taken to avoid retweets. This ensured that a set of unique tweets was obtained. Also, to avoid regional bias, tweets from four major metro cities Mumbai, Bangalore, Kolkata, and Delhi, were collected.
This dataset of 5000 tweets was cleaned by removing punctuations, numbers, URLs, and non-ASCII characters. Stop words were also removed. Latent Dirichlet Allocation method of topic modeling was used to discover the most discussed topics in these tweets. Topics like Taxes, Banking system, Infrastructure development, The Middle Class were found to be at the center of people's conversations. Each tweet from the data set was then manually labeled as one of the topics found by LDA. Naive Bayes classifier was used for the automatic classification of tweets into one of the topics. For training and testing purposes, the data set was split randomly into training and testing sets. The training to testing set was chosen to be in the ratio 3:1. This ratio has been used widely for classification.10 fold cross-validation was performed on the training set. The results could be highly distorted if the training data over fits the test data. To avoid this problem, k-fold cross-validation is used. We have chosen k=10, as is widely used in classification literature (Pennacchiotti, 2011). By choosing k = 10, we divide the dataset into ten equal parts. At one time, 9 of these parts are used for training while the remaining part is used for testing. This process is repeated 10 times so that each part is used for validation once. This is proven to show better accuracy with minimum bias. Also, it solves the problem of over-fitting of data (Kohavi, 1995).

Feature selection

Features are the variables that define the dataset. Feature selection is a way of reducing the number of variables required to define a dataset uniquely. Thus, it is a way of reducing the dimensionality i.e., removing redundant features that do not add to classification accuracy

(Maksoud, 2019 ; Dey, 2018). When the number of variables is large, and the number of training patterns is low, there is a high probability that the classifier works well on the training data but not on the test data. In such cases, over-fitting is said to have occurred (Guyon, 2003). The data needs a large amount of memory and processing time as well because of a large number of variables. Thus, feature selection is important for classification problems. In this paper, we attempt to reduce the number of features by using the words identified by topic modeling as an input to the Document Term Matrix used in the Naive Bayes classifier.

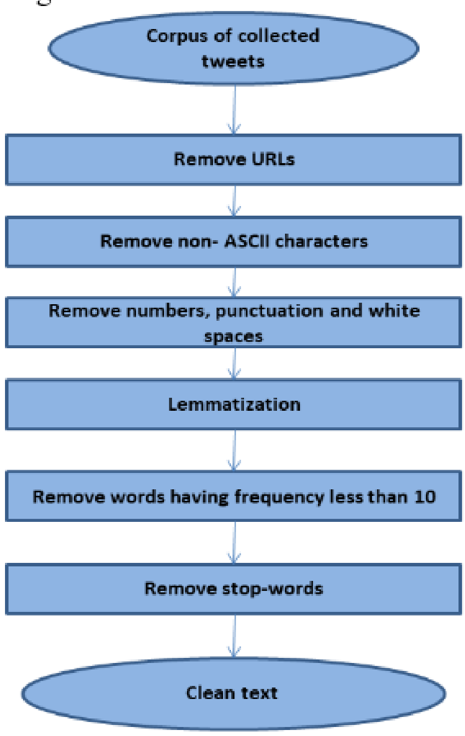
Content Words - As per the Collins dictionary, the content word is a word to which an independent meaning can be given by reference to a world outside any sentence in which the word may occur.

Stopwords - Words like 'the,' 'and' etc. which do not have any meaning on their own are called stopwords. These words are useful for the grammatical relationship between the content words but have little meaning outside the context of the sentence. We use the inbuilt list of English stopwords available in the 'tm' or Text Mining package in R. Besides; we manually list the stopwords which are not included in the dictionary but are present in our corpus. Removing stopwords from the corpus of words in our database reduces the dimensionality of document term matrix.

Less frequent word*s* - We identify words that are less than or equal to two characters in length. Also, words that occur less than ten times in the entire database are identified. These words are removed from the database as they do not contribute to the creation of the document term matrix with words relevant for topic identification.

By removing less frequent words and stopwords, the dimensionality of the document term matrix was reduced from 12436 words to 609 words. The flow is described in figure Ⅰ.

Figure I : Flowchart for cleaning of tweets



Classification Method: Naive Bayes algorithm

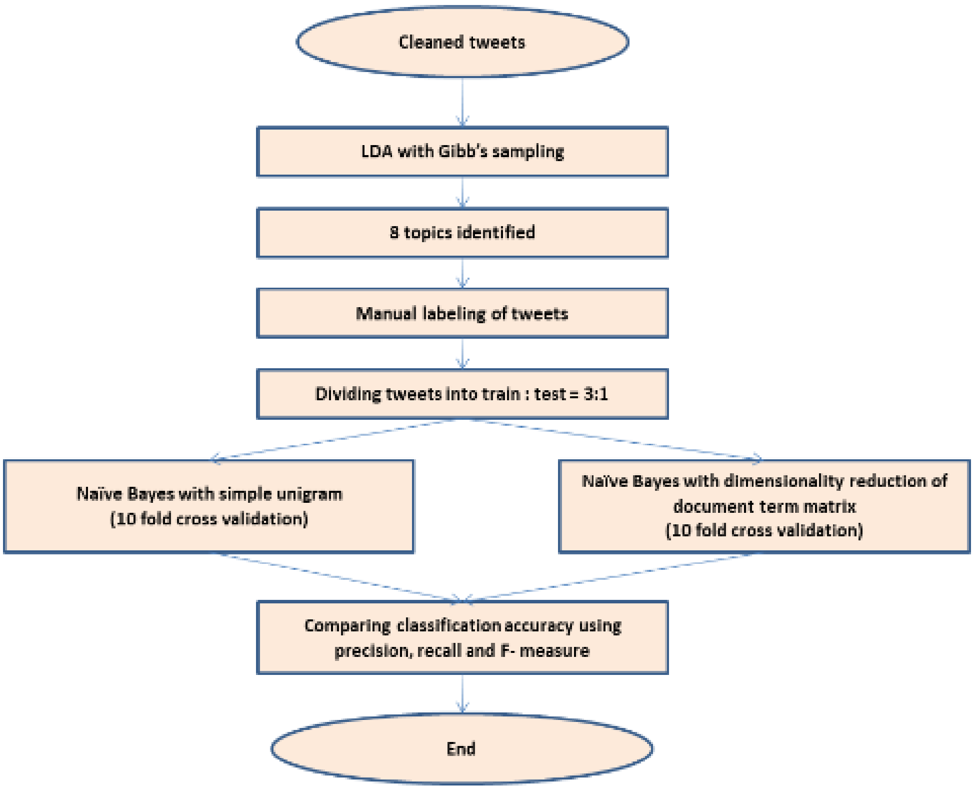
Naive Bayes algorithm is one of the simplest methods used for document classification. It uses conditional probability to predict the possibility of a particular document belonging to a specific class. In our case, we use Naive Bayes classifier to classify the tweets into one of the seven topics identified by topic modeling.

Lexicon based approaches use predefined dictionaries that have words marked against their semantic orientation. Some of the prominent lexicons used for sentiment classification are - SentiWordNet,MPQA subjectivity lexicon, Senticnet etc. Here we use SentiWordNet lexicon. It has three polarities associated with it - positive, negative and neutral. SentiWordNet is one of the most widely used publicly available dictionary for sentiment analysis (Singh , 2016).

Change in sentimen

We divide each topic in test set into two - tweets before budget declaration and tweets after budget declaration (after 5 July 2019). For a particular topic, we find the percentage of positive tweets before budget and after budget. Thus, it is possible to track the change in public mood over the days. The flow is described in figure Ⅱ.

Figure Ⅱ : Flowchart for classification of tweets



Result and Discussion

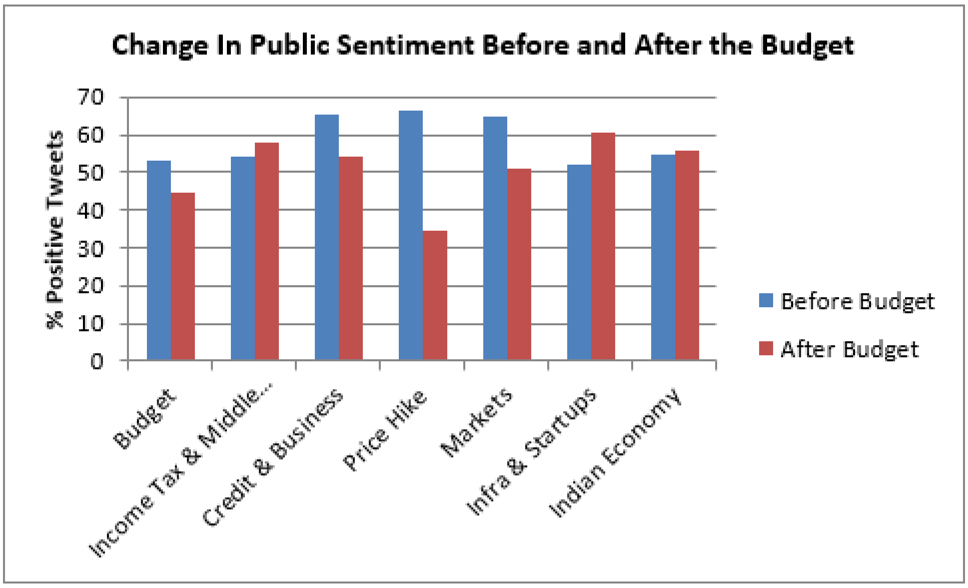
We found that the topics discussed by people were:

- The Budget
- Income tax and Middle Class

- Credit and Business
- Price Hike (Duties, cess)
- Stock Markets
- Infrastructure and Startups
- Indian Economy

It was observed that the excitement about the budget, credit and business and the stock markets has deflated after the announcement of budget. People initially had high expectations about the reforms in credit and banking system. Utilization of tax payers money for bank recapitalization while not taking measures to regulate lending practices to businesses could be the reason for unhappiness among the people. The stock markets also showed a sharp correction after budget with no signs of revival in near future. The differences in the sentiment of the sample used are portrayed in Figure III.

Figure 3: Change in Public sentiment before and after the budget



In this section, we report the classification strength of our model vis-a-vis the simple Naive Bayes classifier using unigram as a feature. We measured the performance of these algorithms using metrics like accuracy, recall, and F- measure. Usually, Precision and recall are used as measures of the effectiveness of the classification algorithm. Accuracy is the percentage of correct classifications. However, when there is a high class-imbalance, the F - measure is a better metric. F-measure combines Precision and Recall.

F- Measure = 2(Precision * Recall)/(Precision + Recall)

Here, accuracy =(TP + TN)/(TP + TN + FP + FN),precision =TP/(TP + FP), recall =TP/(TP + FN)

Where, TP = true positive, TN = true negative, FP =false positive, FN = false negative.

The model was trained on the test set and verified on the test set. The values of precision, recall and F-measure found with our model and the Naive Bayes model are as given in Table III.

Table 3 : Model details

	Precision	Recall	F-measure
Naïve Bayes (Unigram)	0.531	0.495	0.512
Naïve Bayes (with dimensionality reduction)	0.966	0.953	0.959

Conclusion

With the world becoming increasingly data-driven, more and more insights can be drawn from it. This data can be of immense help in governance and public policymaking. Through topic modeling, our paper has presented a way of deciphering the anticipations and expectations of people from the union budget. We also discovered people's reactions to the reforms proposed by the government in the budget. Due to a large amount of data that needs to be processed, ways must be explored that help overcome the limitations on computational power. Our research explores one such method. We were thus successful in reducing the number of features required from 12434 to 609 while improving the F- measure from 0.512 to 0.959. We achieve this by reducing the dimensionality of the document term matrix used in naive Bayes classifier. Besides, the public mood is captured by calculating the percentage of positive tweets in each of the topics identified. It is found that the people were unhappy about the reforms announced for the banking sector. This is reflected in the reduced optimism about the stock markets. Also, the overall sentiment about the budget is found to have taken a beating post budget-announcement. The negative sentiment in these fields possibly mean that the government needs to revisit the policy reforms. Though the policies are aimed at bringing about long term structural changes in the economy, some populist measures are necessary to recover the slump in consumption. A stagnation in the markets may have a cascading effect on other sectors and cause weakening of the economy as a whole. This study supports earlier papers (MICHAEL LAVER, 2003; Singh D. J., 2010; Faricy) that tweets can be used to mine public sentiment on issues of national importance like the budget of a country. This paper also lends credence to earlier works (David M. Blei, 2003; Zhang, 2017; Puranam, 2017) on the usefulness of LDA in identifying topics from tweets. This methodology can successfully help in gauging responses to an event by using social media.

References

1. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, *1*(12), 2009.

2. Lin, D., Matsumoto, Y., & Mihalcea, R. (2011, June). Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.
3. Bansal, B., & Srivastava, S. (2018). On predicting elections with hybrid topic based sentiment analysis of tweets. *Procedia Computer Science*, *135*, 346-353.
4. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.
5. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.
6. Davis, J. A. (1980). Conservative weather in a liberalizing climate: Change in selected NORC general social survey items, 1972–78. *Social Forces*, *58*(4), 1129-1156.
7. Dey, N., Borra, S., Ashour, A. S., & Shi, F. (Eds.). (2018). *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging*. Academic Press.
8. Ladd Jr, E. C., Potter, M., Basilick, L., Dantels, S., & Suszkiw, D. (1979). The polls: Taxing and spending. *Public Opinion Quarterly*, *43*(1), 126-135.
9. Ellis, C., & Faricy, C. (2011). Social policy and public opinion: How the ideological direction of spending influences public mood. *The Journal of Politics*, *73*(4), 1095-1110.
10. Ficcadenti, V., Cerqueti, R., & Ausloos, M. (2019). A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches. *Expert Systems with Applications*, *123*, 127-142.
11. Wei, W., & Gulla, J. A. (2010, July). Sentiment learning on product reviews via sentiment ontology tree. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 404-413).
12. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.
13. Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20-38.
14. Jacoby, W. G. (2000). Issue framing and public opinion on government spending. *American Journal of Political Science*, 750-767.
15. Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).
16. Kwon, N., Shulman, S. W., & Hovy, E. (2006, May). Multidimensional text analysis for eRulemaking. In *Proceedings of the 2006 international conference on Digital government research* (pp. 157-166).Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers ©2012.
17. Maksoud, E. A. A., Barakat, S., & Elmogy, M. (2019). Medical images analysis based on multilabel classification. In *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging* (pp. 209-245). Academic Press.Marco Pennacchiotti, A.-M. P. (2011). A Machine Learning Approach to Twitter User Classification. Fifth International AAAI Conference on Weblogs and Social Media.
18. Rodríguez-Pose, A., & Maslauskaite, K. (2012). Can policy make us happier? Individual characteristics, socio-economic factors and life satisfaction in Central and Eastern Europe. *Cambridge Journal of Regions, Economy and Society*, *5*(1), 77-96.

19. Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American political science review*, *97*(2), 311-331.
20. Bansal, M., Cardie, C., & Lee, L. (2008, August). The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In *COLING 2008: Companion Volume: Posters* (pp. 15-18).
21. Pang, B., & Lee, L. (2009). Opinion mining and sentiment analysis. *Comput. Linguist*, *35*(2), 311-312.
22. Puranam, D., Narayan, V., & Kadiyali, V. (2017). The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors. *Marketing Science*, *36*(5), 726-746
23. Ram, R. (2009). Government spending and happiness of the population: additional evidence from large cross-country samples. *Public Choice*, *138*(3-4), 483-490.
24. Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229-247.
25. Singh, T., & Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science*, *89*(Supplement C), 549-554.
26. Zhang, Y., Moe, W. W., & Schweidel, D. A. (2017). Modeling the role of message content and influencers in social media rebroadcasting. *International Journal of Research in Marketing*, *34*(1), 100-119.