A Naïve Bays Classifier to Classify Movie Review

Neha Raghuvanshi, Prof. J.M. Patil

Abstract— **This paper presents a naïve bays approach to solve sentiment analysis problem. This technique is used to classify movie review in positive or negative domain. To mine the opinion on the web, it is essential to perform a well defined task, which helps us to retrieve the information from the available data on the web. We have started our discussion with the introduction on sentiment analysis, which gives us a insight into sentiment analysis. The detail discussion on various methods proposed by different researchers is also presented. The proposed method talks about the working of naïve bays classifier and is proven to be good method in sentiment analysis domain.**

Index Terms— **Sentiment analysis, Opinion based mining, Naïve bayes classifier.**

I. Introduction

Textual information on the internet is growing every day. To mine and search the textual information is becoming more difficult day by day. The text is prevalent data format on the web, since it is easy to generate and publish. But extracting the information in the proper context from the vast ocean of content is more difficult. The manual efforts are beyond the human control as it takes more time. Therefore, the research problem of automatic categorization and organizing data is apparent. Textual information can be divided into two main domains: facts and opinions. While facts focus on objective data transmission, the opinions express the sentiment of their authors. Initially, the research has mostly focused on the categorization of the factual data. Today, we have web search engines which enable search based on the keywords that describe the topic of the text. The search for one keyword can return a large number of pages. For example, Google search for the word "human" finds more than 2.3 million pages. In recent years, we also contribute our opinion to large number of websites that expresses our views. Opinion can be expressed in different forms. One example may be web sites for reviewing products, such as Amazon, or movie review sites such as Rotten Tomatoes which enable rating of products, usually on some fixed scale as well as leaving personal reviews. These reviews tend to be longer, usually consisting of a few paragraphs of text. With respect to their length and comprehensiveness they tend to resemble blog messages. Other type of web sites contains prevalently short comments, like status messages on social networks like Twitter or article reviews on Digg. Another way of expressing the popularity is to put a rating the popularity of the messages, which can be related to the opinion expressed by the author.

Sentiment Analysis involves determining the evaluative nature of a piece of text. For example, a product review can express a positive, negative, or neutral sentiment (or polarity). Automatically identifying sentiment expressed in text has a number of applications, including tracking sentiment towards products, movies, politicians, etc., improving customer relation models, detecting happiness and well-being, and improving automatic dialogue systems. Over the past decade, there has been a substantial growth in the use of micro blogging services such as Twitter and access to mobile phones world-wide. Thus, there is tremendous interest in sentiment analysis of short informal texts, such as tweets and SMS messages, across a variety of domains (e.g., commerce, health, military intelligence, and disaster management). Sentiment analysis aims to uncover the attitude of the author on a particular topic from the written text. Other terms used to denote this research area include "opinion mining" and "subjectivity detection". It uses natural language processing and machine learning techniques to find statistical and/or linguistic patterns in the text that reveal attitudes. It has gained popularity in recent years due to its immediate applicability in business environment, such as summarizing feedback from the product reviews, discovering collaborative recommendations, or assisting in election campaigns.

Creating systems that can process subjective information effectively requires overcoming a number of novel challenges. To illustrate some of these challenges, let us consider the concrete example of what building an opinion- or review-search application could involve. As we have discussed, such an application would fill an important and prevalent information need, whether one restricts attention to blog search or considers the more general types of search that have been described above. The development of a complete review- or opinion-search application might involve attacking each of the following problems.

(1) If the application is integrated into a general-purpose search engine, then one would need to determine whether the user is in fact looking for subjective material. This may or may not be a difficult problem in and of itself: perhaps queries of this type will tend to contain indicator terms like "review", "reviews", or "opinions", or perhaps the application would provide a "checkbox" to

the user so that he or she could indicate directly that reviews are what is desired; but in general, query classification is a

Manuscript received April, 2016.

Neha Raghuvanshi, PG Student, Dept of CSE, SSGMCE, Shegaon

Prof. J.M. Patil, Assistant Professor, Dept of CSE, SSGMCE, Shegaon

difficult problem — indeed, it was the subject of the 2005 KDD Cup challenge.

(2) Besides the still-open problem of determining which documents are topically relevant to an opinion-oriented query, an additional challenge we face in our new setting is simultaneously or subsequently determining which documents or portions of documents contain review-like or opinionated material. Sometimes this is relatively easy, as in texts fetched from review aggregation sites in which review-oriented information is presented in relatively stereotyped format: examples include Epinions.com and Amazon.com. However, blogs also notoriously contain quite a bit of subjective content and thus are another obvious place to look (and are more relevant than shopping sites for queries that concern politics, people, or other non-products), but the desired material within blogs can vary quite widely in content, style, presentation, and even level of grammaticality.

(3) Once one has target documents in hand, one is still faced with the problem of identifying the overall sentiment expressed by these documents and/or the specific opinions regarding particular features or aspects of the items or topics in question, as necessary. Again, while some sites make this kind of extraction easier—for instance, user reviews posted to Yahoo! Movies must specify grades for pre-defined sets of characteristics of films — more free-form text can be much harder for computers to analyze, and indeed can pose additional challenges; for example, if quotations are included in a newspaper article, care must be taken to attribute the views expressed in each quotation to the correct entity.

(4) Finally, the system needs to present the sentiment information it has garnered in some reasonable summary fashion. This can involve some or all of the following actions: (a) aggregation of "votes" that may be registered on different scales (e.g., one reviewer uses a star system, but another uses letter grades) (b) selective highlighting of some opinions (c) representation of points of disagreement and points of consensus (d) identification of communities of opinion holders (e) accounting for different levels of authority among opinion holders. Note that it might be more appropriate to produce a visualization of sentiment data rather than a textual summary of it, whereas textual summaries are what is usually created in standard topic based multi-document summarization.

II. Literature Review

The research community has studied almost all main aspects of the problem. The most well studied sub problem is opinion orientation classification (i.e., at the document level, sentence level and feature level). In this section, we introduce different methods of sentiment analysis. Kuat Yessenov [1] presents sentiment analysis of movie review comments. They have presented an empirical study of efficacy of machine learning techniques in classifying text messages by semantic meaning. Authors have use movie review comments from popular social network Digg as the data set and classify text by subjectivity/objectivity and negative/positive attitude. Different approaches have been proposed to extract text features such as bag-of-words model, using large movie reviews corpus, restricting to adjectives and adverbs, handling negations, bounding word frequencies by a threshold, and using WordNet synonyms knowledge. The performance is evaluated on accuracy of four machine learning methods - Naive Bayes, Decision Trees, Maximum-Entropy, and K-Means clustering. V.K. Singh, R. Piryani, A. Uddin [2] presents a new Feature-based Heuristic for Aspect-level Sentiment Classification. This paper presents an experimental work on a new kind of domain specific feature-based heuristic for aspect-level sentiment analysis of movie reviews. Authors have devised an aspect oriented scheme that analysis the textual reviews of a movie and assign it a sentiment label on each aspect. The scores on each aspect from multiple reviews arethen aggregated and a net sentiment profile of the movie is generated on all parameters. The SentiWordNet is used which is based scheme with two different linguistic feature selections comprising of adjectives, adverbs and verbs and n-gram feature extraction. They have also used our SentiWordNet scheme to compute the document-level sentiment for each movie reviewed and compared the results with results obtained using Alchemy API. The sentiment profile of a movie is also compared with the document-level sentiment result. The results obtained show that our scheme produces a more accurate and focused sentiment profile than the simple document-level sentiment analysis.

Erik Cambria , Björn Schuller , Yunqing Xia, Catherine Havas [3] presents New Avenues in Opinion Mining and Sentiment Analysis. The findings in this paper are as follows. Gradually, sentiment analysis research is distinguishing itself as a separate field, falling between NLP and natural language understanding. Unlike standard syntactical NLP tasks, such as summarization and auto categorization, opinion mining mainly focuses on semantic inferences and affective information associated with natural language, and doesn't require a deep understanding of text. We envision sentiment analysis research moving toward content-, concept-, and context-based analysis of natural language text, supported by time efficient parsing techniques suitable for big social data analysis. Context-/intent-level analysis ensures the relevance of the opinions gathered. Social context will continue to gain importance, and an intelligent system will have access to the comprehensive personal information of vast numbers of people. Opinion mining will be specific to each user's or group of users' preferences and needs. Opinions won't be generic, but will reflect their source (for example, a relevant circle of friends or users with similar interests, or the selection of a camera for trekking rather than for night shooting). Gilad Katz, Nir Ofek, Bracha Shapira [4] presents Context-based sentiment analysis in 2015. A ConSent model, a novel context-based approach for the task of sentiment analysis is presented. Proposed approach builds on techniques from the field of information retrieval to identify key terms indicative of the existence of sentiment. Authors model these terms and the contexts in which they appear and use them to generate features for supervised learning. The two major strengths of the proposed model are its robustness against noise and the easy addition of features from multiple sources to the feature set. Empirical evaluation over multiple real-world domains demonstrates the merit of proposed approach, compared to state-of the art methods both in noiseless and noisy text. Whilst most researchers focus on assigning sentiments to documents, others focus onmore specific tasks: finding the sentiments of words (Hatzivassiloglou & McKeown 1997), subjective expressions (Wilson et al. 2005; Kim&Hovy

2004), subjective sentences (Pang&Lee 2004) and topics (Yi et al. 2003; Nasukawa & Yi 2003; Hiroshi et al. 2004). These tasks analyze sentiment at a fine-grained level and can be used to improve the effectiveness of a sentiment classification, as shown in Pang & Lee (2004). Instead of carrying out a sentiment classification or an opinion extraction, Choi et al. (2005) focus on extracting the sources of opinions, e.g., the persons or organizations who play a crucial role in influencing other individuals' opinions. Various data sources have been used, ranging from product reviews, customer feedback, the Document Understanding Conference (DUC) corpus, the Multi-Perspective Question Answering (MPQA) corpus and theWall Street Journal (WSJ) corpus. To automate sentiment analysis, different approaches have been applied to predict the sentiments of words, expressions or documents. These are Natural Language Processing (NLP) and pattern-based (Yi et al. 2003; Nasukawa & Yi 2003; Hiroshi et al. 2004; K¨onig & Brill 2006), machine learning algorithms, such as Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM) (Joachims 1998), and unsupervised learning (Turney 2002).

Chen et al. [9] presented a visual analysis system using multiple coordinated views, such as decision trees and terminology variation, to help users to understand the dynamics of conflicting opinions. Wanner et al. [10] described a concise visual encoding scheme to represent attributes, such as the emotional trend of each RSS news item. Both works for analyzing text contents are efficient by using word matching methods. However, they lack semantic analysis. Draper et al. [11] developed an interactive visualization system to allow users to visually construct queries and view the results in real time. For sentiment mining and analysis, Gregory et al. [12] proposed a user-directed sentiment analysis method to visualize aff ective document contents. Although they analyze and visualize emotion, they only use statistical methods. To demonstrate and predict the trend for an event, we suggest that rules about the evolution of public sentiments related to the participants about hot topic types should be modeled and discovered. Collective behavior has many characteristics, such as being spontaneous, zealous, unconventional, and transient. Sentimental contagion and imitation are the main psychological mechanisms of the collective behaviors. Hoyst et al. [13] and Sznajd-Weron et al. [14] proposed two different opinion dynamics models using the aforementioned theory. For example, when discussing a debatable topic on forums, some participants' sentiments can easily be affected by others, which might result in booing or other extreme actions. Several approaches focus on the visual exploration of blogs, forum posts, and Web logs. Adnan et al. [15] used frequent closed patterns to model and analyze data, and create a social network. They also analyzed Web logs by integrating data mining and social network techniques [16]. Indratmo et al. [17] visualized Web tags and comments arranged along a time axis. Dork et al. [18] provided faceted visualization widgets for visual query formulation according to time, place, and tags. Ong et al. [19] proposed an interactive Web-based tree map, News map, to represent the relative number of articles per news item. Fisher et al. [20] found the evolution of topical trends in social media by using line graphs indicating term trends. The aforementioned works focus on social networks, text analysis and knowledge representation of social networks to analyze microblog and forum content without sentiment analysis.

III. Naïve Bays Classifier for movie review

In general sentiment analysis problem is solved with different techniques. Based on the techniques, the general model is as follows.

1. Number of documents are considered
2. The keyword extraction from the reviews
   - ➢ Elimination of the stop words eg. am, is are etc
   - ➢ Identification of the keywords
3. From these keywords, the adjectives and the other opinion oriented keywords are identified
4. Opinion class is identified for each keyword
   - ➢ Number of opinion keywords in each class is identified and based on which the weightage is assigned to each belonging opinion class
   - ➢ Opinion class that will have more number of keywords will be assigned by higher weightage
   - ➢ The aggregative weightage is applied on each opinion class and the collective decision is taken.

Naive Bayes is a simple but effective classification algorithm. The Naive Bayes algorithm is widely used algorithm for document classification.

- – Numberless values taken by a continuous-valued feature
- – Conditional probability often modeled with the normal distribution

$$\hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right)$$

μ_{ji} : mean (avearage) of feature values X_j of examples for which $\mathrm{C} = c_i$

σ_{ji} : standard deviation of feature values X_j of examples for which $C = c_i$

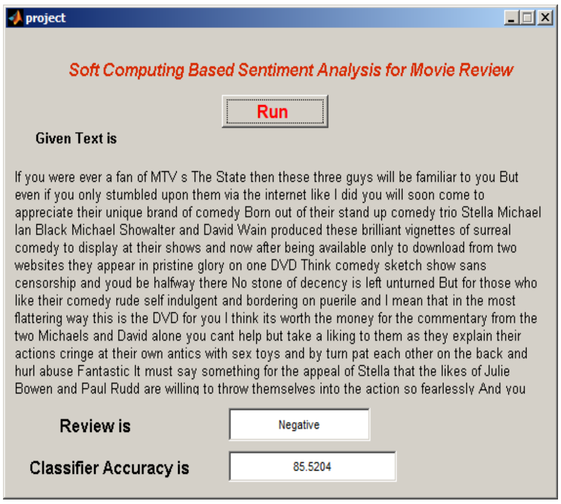
- Learning Phase:
- Output: Normal distributions and
- Test Phase: Given an unknown instance
- Instead of looking-up tables, calculate conditional probabilities with
- all the normal distributions achieved in the learning phrase
- Apply the MAP rule to make a decision

To evaluate the performance of the algorithm its accuracy is calculated based on following formula.

A = (Number of correctly classified documents) / total number of documents

IV. Results

The The algorithm is evaluated on matlab 7.9. Following are the classifier accuracy when it is tested on Large Movie review dataset (IMDB). Following figure shows the effect of proposed algorithm, when it runs. When we click on run, corresponding id is taken and review stored in database is shown in “Given Text is” block. Here the database we use is unlabeled.xml for id=11.



Given Text is

If you were ever a fan of MTV s The State then these three guys will be familiar to you But even if you only stumbled upon them via the internet like I did you will soon come to appreciate their unique brand of comedy Born out of their stand up comedy trio Stella Michael Ian Black Michael Showalter and David Wain produced these brilliant vignettes of surreal comedy to display at their shows and now after being available only to download from two websites they appear in pristine glory on one DVD Think comedy sketch show sans censorship and youd be halfway there No stone of decency is left unturned But for those who like their comedy rude self indulgent and bordering on puerile and I mean that in the most flattering way this is the DVD for you I think its worth the money for the commentary from the two Michaels and David alone you cant help but take a liking to them as they explain their actions cringe at their own antics with sex toys and by turn pat each other on the back and hurl abuse Fantastic It must say something for the appeal of Stella that the likes of Julie Bowen and Paul Rudd are willing to throw themselves into the action so fearlessly And you

Review is Negative

Classifier Accuracy is 85.5204

The classifier accuracy for naïve bays is 85.52 %.

V. Conclusion

This paper gives a a working of naïve bays classifier for categorization of movie review. We have proposed sentiment analysis model based on naïve bayes model. This technique is proposed based on the study of different methods proposed by different researchers. Different types of sentiment analysis model based on word, sentence, feature and document is also discussed and found that document level model is more promising for better results. Result show that accuracy of given classifier is 85.52 which is good enough to perform the sentiment analysis task.

REFERENCES

[1] Kuat Yessenov, Sasa Misailovic,“Sentiment Analysis of Movie Review Comments”, 6.863 Spring 2009 final project, pp.1-17.

[2] Vivek Kumar Singh, Rajesh Piryani, A. Uddin, “Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification”, DOI: 10.1109/iMac4s.2013.6526500, JANUARY 2013.

[3] Erik Cambria , Björn Schuller , Yunqing Xia, Catherine Havasi, “New Avenues in Opinion Mining and Sentiment Analysis”, IEEE Computer Society, March/April 2013, pp. 15-21.

[4] Gilad Katz, Nir Ofek, Bracha Shapira, “ConSent: Context-based sentiment analysis”, Knowledge-Based Systems 84 (2015) 162 ‑ 178.

[5] Hatzivassiloglou, V. & McKeown, K. R. (1997), “Predicting the semantic orientation of adjectives”, In Proceedings of the 8th conference on european chapter of the association for computational linguistics (pp. 174–181). Madrid, Spain.

[6] Hiroshi, K., Tetsuya, N., & Hideo, W. (2004), “Deeper sentiment analysis using machine translation technology”, In Proceedings of the 20th international conference on computational linguistics (COLING 2004), August 23 – 27, 2004 (pp. 494–500). Geneva, Switzerland.

[7] Nasukawa, T. & Yi, J. (2003), “Sentiment analysis: capturing favorability using natural language processing”, In Proceedings of the 2nd international conference on Knowledge capture, October 23–25, 2003. (pp. 70–77). Florida, USA.

[8] Turney, P. D. (2002), “Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews”, In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL), July 6–12 , 2002 (pp. 417–424). Philadelphia, PA, USA.

[9] C. Chen, F. Ibekwe-SanJuan, and E. SanJuan, “Visual analysis of conflicting opinions,” in Proc. IEEE Comput. Soc. Symp. Visual Analytics, Chicago, IL, USA, 2006, pp. 59–66.

[10] F. Wanner, C. Rohrdantz, F. Mansmann, D. Oelke, and D. Keim, “Visual sentiment analysis of RSS news feeds featuring the US presidential election in 2008,” presented at theWorkshop on Visual Interfaces to the Social and the Semantic Web, Sanibel Island, FL, USA, 2008.

[11] G. Draper and R. Riesenfeld, “Who votes for what? A visual query language for opinion data,” IEEE Transa. Vis. Comput. Graphics, vol. 14, no. 6, pp. 1197–1204, Nov./Dec. 2008.

[12] M. Gregory, N. Chinchor, P.Whitney, R. Carter, E. Hetzler, and A. Turner, “User-directed sentiment analysis: Visualizing the affective content of documents,” in Proc.Workshop Sentiment Subjectivity Text, 2006, pp. 23– 30.

[13] J. Hoyst, K. Kaceperski, and F. Schweitzer, Annual Reviews of Computational Physics IX. Singapore:World Scientific, 2001.

[14] K. Sznajd-Weron, “Sznajd model and its applications,” Acta Phys. Polonica B, vol. 36, p. 2537, 2005.

[15] M. Adnan, R. Alhajj, and J. Rokne, “Identifying social communities by frequent pattern mining,” in Proc. 13th Int. Conf. Inf. Vis., 2009, pp. 413–418.

[16] M. Adnan, M. Nagi, K. Kianmehr, M. Ridley, R. Alhajj, and J. Rokne, “Promotingwhere,when and what? An analysis ofweb logs by integrating data mining and social network techniques to guide eCommerce business promotions,” J. Social Netw. Anal. Mining, vol. 1, no. 3, pp. 173–185, 2010.

[17] J. Vassileva and C. Gutwin, “Exploring blog archives with interactive visualization,” in Proc. Conf. Adv. Vis. Interfaces, 2008, pp. 39–46.

[18] M. Dork, S. Carpendale, C. Collins, and C. Williamson, “VisGets: Coordinated visualizations for Web-based information exploration and discovery,” IEEE Trans. Vis. Comput. Graph., vol. 14, no. 6, pp. 1205–1212, Nov./Dec. 2008.

[19] T. Ong, H. Chen,W. Sung, and B. Zhu, “Newsmap: A knowledge map for online news,” Decision Support Syst., vol. 39, no. 4, pp. 583–597, 2005.

[20] D. Fisher, A. Hoff, G. Robertson, and M. Hurst, “Narratives: A visualization to track narrative events as they develop,” in Proc. IEEE Symp. Vis. Anal. Sci. Technol., 2008, pp. 115–122.