International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 6 Issue 2 July - December 2016
International Manuscript ID : 22308849072016-09

Effective Approach for Sentiment Opinion Mining using Natural Language Extraction and Tweets Evaluation

Sukhdeep Kaur
Research Scholar
Department of Computer Science and Engineering
Ludhiana College of Engineering and Technology
Ludhiana, Punjab, India

Ravinder Kamboj
Assistant Professor
Department of Computer Science and Engineering
Ludhiana College of Engineering and Technology
Ludhiana, Punjab, India

ABSTRACT
Sentiment Data Mining and Predictive Analysis refers to the approach in which the live extraction of timeline, tweets and related information from social media is fetched for deep analysis and learning. In this process, a number of libraries and algorithmic approaches are used to evaluate the positive or negative opinions. In this research work, a unique and effective approach for opinion mining is implemented making use of deep learning and extraction. The results show the effective predictive analysis in very less complexity factor in getting the live dataset from Twitter and sentiment score.

Keywords - Sentiment Mining, Opinion Mining, Stanford NLP, Twitter Mining

INTRODUCTION
Big data can be characterized by 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily.

Big data can be characterized by 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily.

Although the demand for big data analytics is high, there is currently a shortage of data scientists and other analysts who have experience working with big data in a distributed, open source environment. In the enterprise, vendors have responded to this shortage by creating Hadoop appliances to help companies take advantage of the semi-structured and unstructured data they own.

Big data can be contrasted with small data, another evolving term that's often used to describe data whose volume and format can be easily used for self-service analytics. A commonly quoted axiom is that "big data is for machines; small data is for people."

2. RELATED WORK

Bifet (2009) [8] - Twitter is a micro-blogging service built to discover what is happening at any moment in time, anywhere in the world. Twitter messages are short, and generated constantly, and well suited for knowledge discovery using data stream mining. The authors briefly discuss the challenges that Twitter data streams pose, focusing on classification problems, and then consider these streams for opinion mining and sentiment analysis. To deal with streaming unbalanced classes, this work proposes a sliding window Kappa statistic for evaluation in time-changing data streams. Using this statistic this work performs a study on Twitter data using learning algorithms for data streams.

Bollen (2009) [2] - The authors in this paper compare the results to the values recorded by stock market and crude oil price indices and major events in media and popular culture, such as the U.S. Presidential Election of November 4, 2008 and Thanksgiving Day. This work finds that events in the social, political, cultural and economic sphere do have a significant, immediate on the various dimensions of public mood. The authors speculate that large scale analyses of mood can provide a solid platform to model collective emotive trends in terms of their predictive value with regards to existing social as well as economic indicators.

Bollen (2010) [1] - This work analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). This paper cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. The results in this paper indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. The authors find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6%.

Davidov (2010) [6] - Automated identification of diverse sentiment types can be beneficial for many NLP systems such as review summarization and public media analysis. In some of these systems there is an option of assigning a sentiment value to a single sentence or a very short text. In this paper the authors proposes a supervised sentiment classification framework which is based on data from Twitter, a popular microblogging service. By utilizing 50 Twitter tags and 15 smileys as sentiment labels, this framework avoids the need for labor intensive manual annotation, allowing identification and classification of diverse sentiment types of short texts. The authors evaluate the contribution of different feature types for sentiment classification and show that the proposed framework successfully identifies sentiment types of untagged sentences. The quality of the sentiment identification was also confirmed by human judges. This paper also explores dependencies and overlap between different sentiment types represented by smileys and Twitter hashtags.

Asur (2010) [3] – Sentiment Analysis is important part for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, the authors demonstrate how social media content can be used to predict real-world outcomes. In particular, this work uses the chatter from Twitter.com to forecast box-office revenues for movies. This paper shows that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. This work further demonstrates how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media.

Tan (2011) [4] – The authors show that information about social relationships can be used to improve user-level sentiment analysis. The main motivation behind the approach is that users that are somehow "connected" may be more likely to hold similar opinions; therefore, relationship information can complement what we can extract about a user's viewpoints from their utterances. Employing Twitter as a source for our experimental data, and working within a semi-supervised framework, we propose models that are induced either from the Twitter follower/followee network or from the network in Twitter formed by users referring to each other using "@" mentions. The proposed transductive learning results reveal that incorporating social-network information can indeed lead to statistically significant sentiment classification improvements over the performance of an approach based on Support Vector Machines having access only to textual features.

Saif (2012) [5] - Sentiment analysis over Twitter offer organisations a fast and effective way to monitor the publics' feelings towards their brand, business, directors, etc. A wide range of features and methods for training sentiment classifiers for Twitter datasets have been researched in recent years with varying results. In this paper, we introduce a novel approach of adding semantics as additional features into the training set for

sentiment analysis. For each extracted entity (e.g. iPhone) from tweets, we add its semantic concept (e.g. "Apple product") as an additional feature, and measure the correlation of the representative concept with negative/positive sentiment. The authors apply this approach to predict sentiment for three different Twitter datasets. The results show an average increase of F harmonic accuracy score for identifying both negative and positive sentiment of around 6.5% and 4.8% over the baselines of unigrams and part-of-speech features respectively. We also compare against an approach based on sentiment-bearing topic analysis, and find that semantic features produce better Recall and F score when classifying negative sentiment, and better Precision with lower Recall and F score in positive sentiment classification.

Saif (2012) [7] - Twitter has brought much attention recently as a hot research topic in the domain of sentiment analysis. Training sentiment classifiers from tweets data often faces the data sparsity problem partly due to the large variety of short and irregular forms introduced to tweets because of the 140-character limit. In this work the authors proposes using two different sets of features to alleviate the data sparseness problem. One is the semantic feature set where this work extracts semantically hidden concepts from tweets and then incorporate them into classifier training through interpolation. Another is the sentiment-topic feature set where we extract latent topics and the associated topic sentiment from tweets, then augment the original feature space with these sentiment-topics. Experimental results on the Stanford Twitter Sentiment Dataset show that both feature sets outperform the baseline model using unigrams only. Moreover, using semantic features rivals the previously reported best result. Using sentiment topic features achieves 86.3% sentiment classification accuracy, which outperforms existing approaches.

3. PROPOSED WORK

There is need to propose and work out a novel approach for social media mining with the predictive analysis of the results and relate popularity score

Research Objectives

- To fetch the live timeline and real time data from social media
- To identify the social media platform with the access to API so that real time data can be fetched out.
- To perform the predictive analysis of the sentiment or opinion score of tweets individually and cumulative phase.

4. IMPLEMENTATION

1. Development of an effective GUI on Java Platform to fetch the live tweets from Twitter
2. Real Time connectivity of Java - Twitter
3. Implementation of Data Mining Algorithms
    a. Sentiment Mining
    b. Predictive Analysis

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 6 Issue 2 July - December 2016
International Manuscript ID : 22308849072016-09

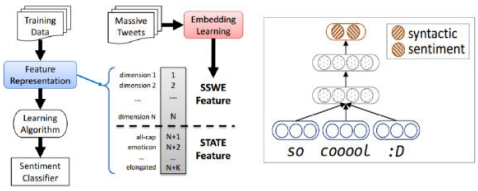


Figure 1 – Sentiment Score Analytics

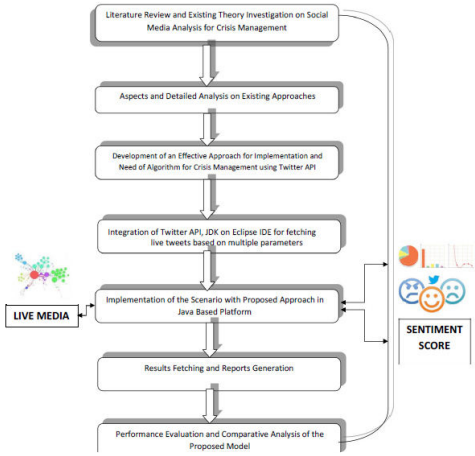


Figure 2 –Flow of Work

Figure 3 –Procedural Dimensions

RESULTS

Following tools and technologies are used for implementation and fetching of results -

- Apache Tomcat
- Java
- Twitter APIs
- AJAX
- Twitter4J
- MySQL
- Notepad++
- Eclipse IDE
- JSON (JavaScript Object Notation)

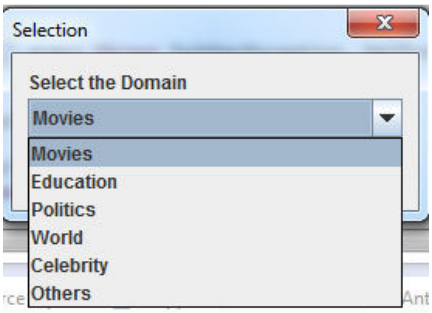


Figure 4 –Selection of Domain for Fetching

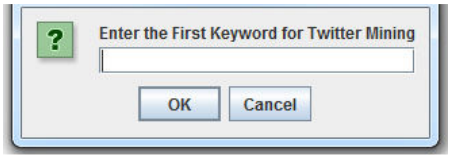


Figure 5 – Selection of First Keyword from Domain

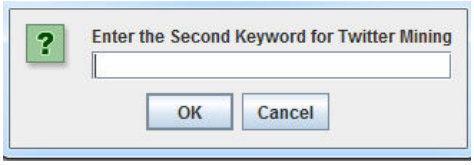
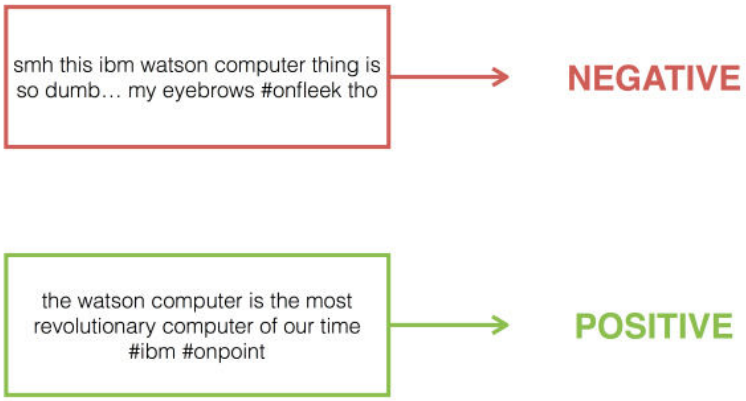


Figure 6 – Selection of Second Keyword for Twitter Mining

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 6 Issue 2 July - December 2016
International Manuscript ID : 22308849072016-09

Table 1 – Words in the Sentiment Text and Scoring Aspects

Text	Base Common Words Layer	Emoticon Layer	Final Score
I **love** dogs. They are **wonderful!** 😍 😍	*Positive words:* love,1 | wonderful!,5 *Negative words:* ***Layer score:*** 2	*Positive emoticons:* 😍 | 😍 *Negative emoticons:* ***Layer score:*** 2	***4***
I **hate** brussel sprouts. They are **terrible.** 💩	*Positive words:* *Negative words:* hate,1 | terrible.,6 ***Layer score:*** -2	*Positive emoticons:* *Negative emoticons:* 💩 ***Layer score:*** -1	***-3***
This is **great!** But also **bad.**	*Positive words:* great!,2 *Negative words:* bad.,5 ***Layer score:*** 0	*Positive emoticons:* *Negative emoticons:* ***Layer score:*** 0	***0***



Flow of the Training and Predictive Analysis

```
TRAINMULTINOMIALNB(C, D)
 1  V ← EXTRACTVOCABULARY(D)
 2  N ← COUNTDOCS(D)
 3  for each c ∈ C
 4  do N_c ← COUNTDOCSINCLASS(D, c)
 5     prior[c] ← N_c/N
 6     text_c ← CONCATENATETEXTOFALLDOCSINCLASS(D, c)
 7     for each t ∈ V
 8     do T_ct ← COUNTTOKENSOFTERM(text_c, t)
 9     for each t ∈ V
10     do condprob[t][c] ← (T_ct+1) / Σ_t'(T_ct'+1)
11  return V, prior, condprob

APPLYMULTINOMIALNB(C, V, prior, condprob, d)
 1  W ← EXTRACTTOKENSFROMDOC(V, d)
 2  for each c ∈ C
 3  do score[c] ← log prior[c]
 4     for each t ∈ W
 5     do score[c] += log condprob[t][c]
 6  return arg max_c∈C score[c]
```

Output Format and Sentiment Score
[Total Sentiment Score -> 25]
[POSITIVE Sentiment Score -> 15]
[NEGATIVE Sentiment Score -> 0]
---------->
[Cumulative Sentiment Score -> 25]
Simulation Scenario Execution Time in MilliSeconds => 76
Sentiment Score : First Keyword -> china
18
Percentage Popularity : First Keyword -> 41.860466
Sentiment Score : Second Keyword -> india
25
Percentage Popularity : Second Keyword -> 58.139534

CONCLUSION AND FUTURE WORK

This work focus on the live extraction of social media for the predictive analysis and popularity score of the real world objects. This proposed implementation is effective in terms of less execution time and complexity. This work can be integrated with the metaheuristic or hyperheuristic approaches for further enhancement and accuracy level.

REFERENCES

[1] Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*, *52*(1), 5-19.

[2] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1-8.

[3] Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *arXiv preprint arXiv:0911.1583*.

[4] Asur, S., & Huberman, B. (2010, August). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492-499). IEEE.

[5] Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P. (2011, August). User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1397-1405). ACM.

[6] Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012* (pp. 508-524). Springer Berlin Heidelberg.

[7] Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 241-249). Association for Computational Linguistics.

[8] Saif, H., He, Y., & Alani, H. (2012). Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org).

ISSN (Online) : 2230-8849
Volume 6 Issue 2 July - December 2016
International Manuscript ID : 22308849072016-09

[9] Bifet, A., & Frank, E. (2010, January). Sentiment knowledge discovery in twitter streaming data. In *Discovery Science* (pp. 1-15). Springer Berlin Heidelberg.