Opinion Mining with Real Time Ontology Streaming Data

Dr.G. Vithya, J. Naren and V. Varun

Abstract--- *Social Networking, the fastest mode of finding individuals with heterogeneous opinions on various issues is the current trend in today's world. There are many social networking sites like facebook, twitter etc where not only information exchange but also sharing opinions happens. Sentiment Analysis or Opinion Mining deals with various emotions and its analysis as positive, negative and neutral due to the mood of a particular individual. The work ultimately focuses on system built with opinions mined from data extracted live from Twitter. The development in a particular field could be efficiently analyzed based on Opinion Mining. Extraction of major important features is done with Ontologies and its analysis with feature quotient. In the proposed work all attributes are analyzed and individual scores are allotted.*

Keywords--- *Sentiment Analysis, Semantic Web, Data Mining, Prediction, Ontology.*

I. INTRODUCTION

Social media is basically a Collaborative resource sharing platform, which functions as a Web 2.0 application. It enables the user to share views and communicate over a common networking platform. Here the user will be able to create, co-create, discuss and share the contents. Twitter, Facebook, Instagram are some of the popular social media sites. Social media mining is the method of analyzing and extracting useful patterns from social media contents. It helps market researchers to study the view, trend and depth of the Market. It encompasses the tools and methods used to define measure, analyze and extract such patterns from large amount of social media contents available.

Real time sentimental analysis of twitter is one such social media mining process, which deals with the usage of user generated contents and extracts useful patterns relative to the subject. Twitter is one such social networking platform that enables users to send and read short messages of 140 characters called tweets. Tweets used here are the user's view and perspective carrier texts on subject of interests.

Hash-tag is another important characteristic feature of twitter, which corresponds to the tweet's subject. Usage of hash-tags in tweets helps in the classification based on subject. Through hash-tag search all the tweets that are associated with that hash-tag are sorted which help data gathering and extraction based on the subject. The real time sentiment analysis system takes tweets as input to study sentiment perspectives related to the subject. Real Time Ontology is an Ontology which is constructed for streaming A Real Time Ontology, Ontology for the Streaming data, is constructed to study the interrelationships between the tweets. The classification of tweets based on the subject is done through hash-tag search. The categorized tweets are then retrieved from twitter to be fed in to the system and can be done through the usage of twitter API crawlers.

Dr.G. Vithya, Professor, School of Computing, KL University, Vijayawada, AP, India. E-mail: vithyamtech@gmail.com
J. Naren, Assistant Professor, School of Computing, SASTRA Deemed University Thanjavur, Tamil Nadu, India.
E-mail: naren.jeeva3@gmail.com
V. Varun, B.Tech Computer Science and Engineering, SASTRA Deemed University, Thanjavur, Tamil Nadu, India.

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019

The process involves three main stages, Data pre- processing, Feature Extraction and Sentimental Classification. Data pre-processing corresponds to retrieval of tweets which forms the system's input entity. The tweets as such cannot be fed into the system directly, as it may contain irrelevant entities. Pre-processing involves stripping away retweets, conversion of tweets into data frames and corpus, extraction of re-tweet count, Removal of punctuations, numbers and URLs. Ontology based approach aims to deliver every key attribute a score, which can be aggregated to determine the overall score that describes the sentiment perspective of the tweets. (For e.g. Music, Script, Graphics are the attributes of the subject, Movie). Sentimental classification process involves measuring the polarity of the feature elements identified from tweets. Individual polarity ratings are calculated for each feature by identifying the positive and negative terms associated with them. Finally, the values are delivered to provide visualizations in a dynamic environment through plot graphs, histograms et al.

II. PROBLEM DEFINITION

Sentiment analysis is a type of Natural language processing method to analyze and deliver the statistical data about people's perception over subject of interest. Due to wide spread usage of social media, sentimental analysis of social media plays a key role in analyzing the market reach and characteristic aspect of the product. Twitter sentimental analysis is one such system gaining popularity in recent times. The objectives of the twitter sentimental analysis system are,

- Classification of tweets based on the subject using hashtags search.
- Real time tweets retrieval using twitter API crawler.
- Data preprocessing methods to convert the rawdata into system readable form.
- Establishing the entity relationship of each of the characteristic attribute of thesubject.
- Providing sentimental score for each of these attributes based on the functional influence of the attribute over the subject.
- Calculating the overall sentimental score, this corresponds to the aggregate of each attribute's sentimental score.
- Knowledge representation through heat maps, bar plot etc.

III. LITERATURE SURVEY

Existing system

Base paper: "Contextual semantics for sentiment analysis of twitter" by Hassan Saif, Yulan he, Miriam Fernandez, Harith alan. Knowledge Media Institute, The Open University, United Kingdom, 2016.

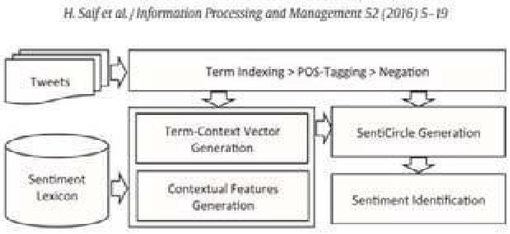

H. Saif et al. / Information Processing and Management 52 (2016) 5–19
Tweets
Term Indexing > POS-Tagging > Negation
Term-Context Vector Generation
SentiCircle Generation
Sentiment Lexicon
Contextual Features Generation
Sentiment Identification

Fig. 1.3.1: Systematic workflow of senticircle

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019

Contextual semantics in sentimental analysis of twitter presents semicircles, a lexicon based approach. Semicircles consider the patterns of words along with contextual position of those patterns to evaluate precise meaning expressed through that pattern of words. The method operates at both entity level and tweet level. The output is the senticircle representation which expresses sentimental perspective of the statement. Based on the region in which a particular word falls, it is categorized into very positive, positive, very negative and negative terms. Overall sentimental score can be calculated from the aggregates derived from senti-circles.

The twitter datasets used in the papers are OMD Diakopoulos and shamma (2010), HCM Speriosu et al (2011), STS-Gold Saif et al (2013)

Table 1.3.1: Twitter dataset

Dataset	**Tweets**	**Positive**	**Negative**
OMD	1081	393	688
HCR	1354	397	957
STS-Gold	2034	632	1402

Improvements made from the existing system

1. Subject oriented classification of tweets through hashtags search.
2. Ontology based approach, which aids in establishing entity-relationships efficiently.
3. Feature extraction process which aims to deliver more details about each and every attribute associated with the subject.
4. Individual Sentiscore for each attribute to define their feature score.
5. Overall score is calculated as an aggregate of the individual Sentiscore of their attributes and influence of the attribute over subject (for e.g. the scope of humor score doesn't mean much in an actual action/thriller movie.

Related work

Natural language processing has different branches of studies dealing with different functional models and methodologies. Sentimental analysis is one such process that deals with the text classification in order to determine the perspective of the author. That could either be a positive or negative like, the application of navies based, support vector machine and classification algorithm. With the help of the algorithm sentimental reviews are categorized. [1]

Feature based opinion mining methodologies deals with the establishment of concepts present in feature and advanced mathematical method in the sentimental analysis process. The target of the proposed methodologies is to develop a method which makes use of ontology in feature based opinion mining at the feature selection stage and to develop an innovative vector analysis method for the sentimental analysis. The promising results inferred from the real world theme scenarios prove the effectiveness of the proposed system. [2]

The paper deals with a combinational approach that integrates rule based classification supervised learning and machine learning. The real world scenarios of movie reviews, reviews on product and social media are tested using the method. And results show improved efficiency in calculations as values of f1at both micro and macro levels, whereas f1 is a measure of both precession and recall aspect. [3]

Web 2.0 has becomes more popular by the way users using the internet, by improved opinion sharing. Twitter is one of the Web 2.0 based application which is used by users to sharing opinions on a product. So, twitter gives a rich opinion sharing data which is analyzed to get the overall score of subjects. Sentiment analysis of the text becomes inefficient due to character limit provided by twitter. In order to increase the efficiency, Ontology based approach is proposed. The main aim of the approach is to find individual sentimental score of distinct features in tweets. [4]

Due to the rapid increase in the usage of social media, analysis of social media content plays the vital role in analyzing the market’s depth. Twitter is one such social media where the user posts real time opinions about topics of interest. The paper introduces the methodology which is hybrid approach that incorporates both corpus method and dictionary method to determine the opinion words orientation present in tweets. [5]

The paper introduces novel machine learning method which identifies the subjective portions and applying text categorization technique over it. To find minimum cuts in the graph, the extraction of subjective portion technique can be employed which helps in the condition of cross sentence contextual constraints. [6]

The analysis of the twitter sentiment is one of the fast and efficient ways to study the public’s feeling towards subject of interest. So, the paper introduces the novel method of semantic approach in sentimental analysis. In the method Semantics is added to each and every feature extracted and correlation of the representative concept is measured by positive/negative sentiment. The approach shows significant increase of f harmony accuracy score around 6.5% for positive and negative sentiment analysis. And the method is proved to have better precision with lower recall and f score than the sentiment bearing topic analysis. [7]

In the paper linguistic feature of texts are utilized to detect the sentiment of the twitter context. Lexical resources and features about informal and creative languages which are used in micro-blogging are utilized in the approach. The method takes supervised approach to the existing problem but makes use of the existing hashtags to develop training dataset.

[8]. The paper deals with the methodology of ontology based sentimental classification to study and analyze online reviews of the users. Here implementation and experimentation deals with the support vector machine text classification approach that is based on the lexical variable ontology. Test results prove the effectiveness of the method in sentimental classification of text content. [9]

IV. PROPOSED ARCHITECTURE

Conceptual system model

Mining of sentiment or opinion words, analysis of opinion words and arguments in the text corresponds to sentiment analysis. The clustering process corresponds to opinion and sentiment detail clusters about the object. The paper deals with the method of ontology based sentiment clustering to cluster study and analyze the reviews. Domain ontology method is proposed in the method to extract the related generic category which helps in identification of the class which the subject falls.

FCM clustering process is incorporated with the knowledge derived from the domain ontology and relative adjectives by which the attribute based fuzzy score is calculated, where the FCM corresponds to fuzzy c-clustering. [10]. The computational treatment of opinions, subjectivity and sentiment of the texts corresponds to sentiment analysis (SA). The survey gives a comprehensive overview of the updates that sentimental analysis has seen over the years. The paper proposes many algorithmic advancements and where SA applications are. The algorithms used are of different methodologies and different data scopes. Many type of sentimental classification techniques such as supervised learning, lexicon based approach, formal concept analysis are explained and the efficiency is compared and presented. [11]. The Proposed work done is different from other works when it comes to Sentiments gathered from real time streaming data and an Ontology introduced in the approach helps here to study the interrelationships between tweets.

A

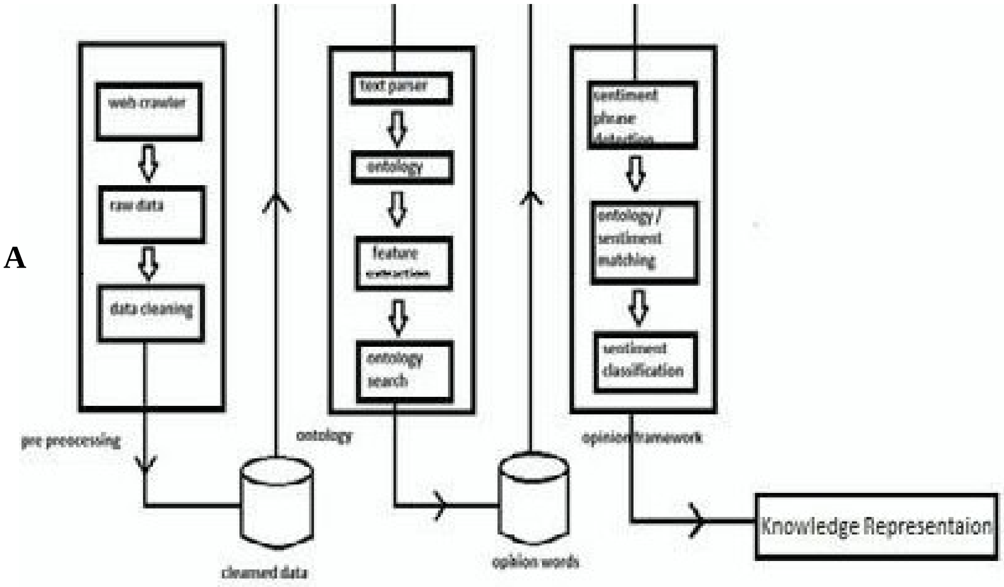


Figure 1.4.1: Conceptual Model of the System The process flow depicts how data from twitter are cleaned and the feature is extracted, how the sentiment is classified and the Sentiscore is represented.

Raw data input

With the help of twitter retrieval API crawler tweets are collected with respect to the user specified subject. The tweets are input from which the sentiments or opinion words are extracted. The keywords related to the components of the module are,

Twitter API crawler. Access token

- Secret key.

Preprocessing

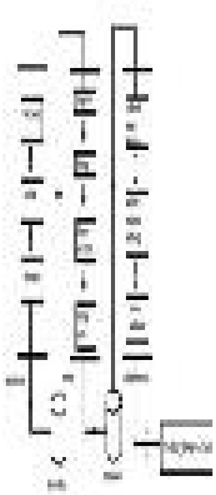
Tweets that are retrieved by using twitter API crawler is cleaned in order to remove the unnecessary symbols, numbers and URLs except full stops. The cleaned data is given to the ontological unit. The keywords related to the components of the module are,

- Stripping away retweets.

- Conversion of tweets into data frames and corpus.
- Extraction of re-tweet count.
- Removal of punctuations, numbers and URLs

Ontology

Ontology represents types, properties and the inter- relationship of the entities that exists for particular domain of discourse. Ontological unit comprises the domain feature categorization and extraction. In order to find the feature, cleaned tweets are taken as the input and the stops words like ‘the’, ’is’, ’etc’ are removed. The key components presents in the process are,



- Text parser.
- Ontology Detection.
- Feature Extraction.
- Ontology search

Opinion Framework

Here the output of ontology is taken as the input and it is classified at individual entity level to determine the overall sentiment expressed through that sentence. Determination of positive or negative words through identification of keywords with the positive and negative words data is done. Assigning true or false based on their polarity is established. The individual Score is aggregated to determine the overall score. The process steps of the opinion framework module are,

- Sentiment phrase detection.
- Ontology/Sentiment detection.
- Sentiment classification
- Pie chart.
- Histograms.
- Plot graphs
- Bar graphs, etc

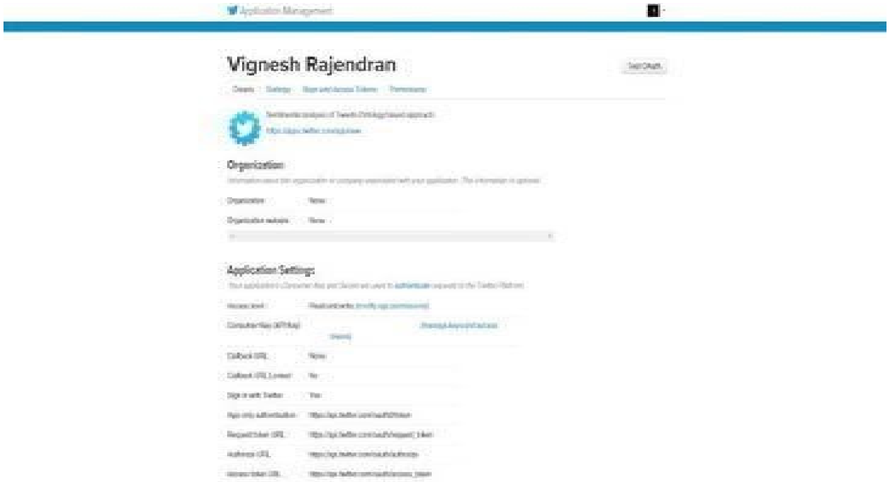
Methodology and Approach Twitter API crawler

Open Authentication (OAuth) is an open standard for confirmation that is embraced by Twitter to give access to the secured data. OAuth gives a more secure distinct option for customary confirmation approaches utilizing a three-way handshake. Here is the reference for more insights about OAuth: Twitter OAuth.

The validation of API asks for on Twitter is done through OAuth. Take note of that Twitter APIs must be gotten to by enlisted applications (e.g., the crawlers you will create in this task). With a specific end goal to enroll the application, initially Twitter account is needed. After that, one has to tie the Twitter account with the application enrolled (i.e., crawlers). When the coupling prepare is completed, the keys and tokens are got (i.e., a couple of purchaser key and shopper mystery and a couple of get to token and get to token mystery) for the application.

Here are the principle ventures for the above enrollment and tying handle:

- Enroll the application to Twitter and get the customer keys.
- Go to https://dev.twitter.com/applications/new and enroll another application to twitter for the task. A name at one’s decision for the application can be picked. For Website URL, you can either utilize your own particular landing page or essentially sort
- “http://” IP address of the machine.
- Fill in every single required field, acknowledge the Developer Agreement, settle the CAPTCHA and present.
- Acquire the shopper key (API key) and buyer mystery from the screen and utilize them in the application (i.e., crawlers).
- Tie the Twitter record and application and get the get to tokens.
- In the website page of the application, tap the Keys and Access Tokens tab, then look down and click create token.
- In the site page of the application, tap the Permissions tab and design an application with the authorization level one requires (in particular, read- compose with-direct messages).



Application Management

Vignesh Rajendran

Organization

Application Settings

Figure 1.4.2: Setting up Twitter API crawler

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019

Get the showed to get token and get to secret token from the screen and utilize them in the application.

Vignesh Rajendran

Organization

Application Settings

Fig. 1.4.3: Authorization screen for Twitter API crawler

Pre-processing

Information preprocessing is an information mining method that includes changing crude information into a reasonable arrangement. Genuine information is frequently fragmented, conflicting, and/or ailing in specific practices or inclines, and is prone to contain numerous mistakes. Information preprocessing is a demonstrated technique for determining such issues. Information preprocessing gets ready crude information for further handling. The process of data pre-processing involves,

Extracting tweets

Subject is provided for sentiment calculation is given to the function, Search Twitter. The function extracts tweets from the twitter, which is used to calculate the sentiment. Stripping away retweets:

- Extracted tweets contain retweets (which is like sharing the other users tweets). Analyzing the retweets takes lot of time, which reduces efficiency of the program. So the retweets are stripped away and counting of the stripped tweets is calculated. Counting is to provide the score of those retweets.

Converting the tweets into data Frame:

- Tweets that are extracted from twitter will be stored in the format which is uneven. So, the tweets are converted into data frames, which store the tweets in the form of data tables. The value representation accounts for value extraction efficiency.
- Extracting the re-tweet count:
- Extracting the number of tweet counts, in turn reduce the redundancy of tweets analyzed.
- Converting the data frame into a corpus:

Tweets that are converted into data frames are in turn converted to corpus for representing the tweets as a text. The text can be used for Sentimental classification processing of tweets.

- To convert to lower case:

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019

The tweets are then converted to lowercase in order to form coherence units.

- To remove punctuations without removing full- stop: Punctuation is considered to be the garbage processing, since it is uncategorized in the process.

To remove numbers in the corpus:

Numbers are also uncategorized in the process, since it doesn’t prove to make much sense in sentiment extraction.

- Convert the corpus to data frame:

The corpus form of texts are then reverted back to the data frame that renders separation of opinion words which forms the input of sentiment extraction process.

Ontology and domain feature extraction

Ontology represents types, properties and the inter- relationship of the entities that exists for particular domain of discourse.

Ontological unit comprises of

- Domain Feature Categorization.
- Domain Feature extraction. However Data pre-processing is a prerequisite process to render effective functioning of Feature extraction process.

Domain feature categorization

The main reason for the ontology based approach is to extract feature components out of it. For that, the feature has to be identified in order to render the extraction process.

The feature corresponds to the attributes of the system that are influential in analyzing the sentiment aspect of the text. For example the feature functions of the movie are Graphics, Direction, Screenplay, etc.

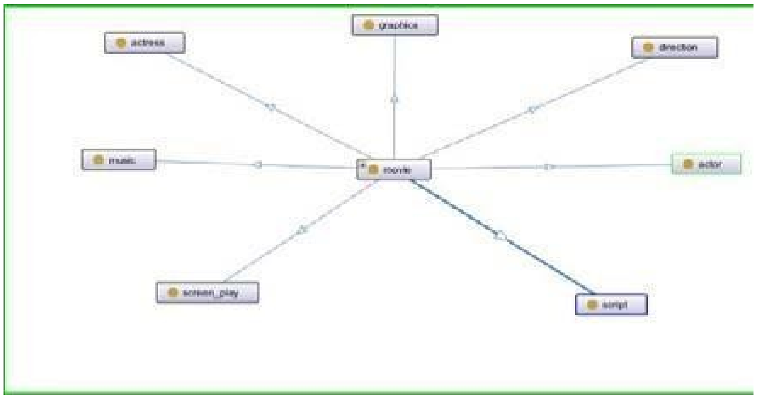


Figure 1.4.2: Domain Feature Categorization for Movie

V. DOMAIN FEATURE EXTRACTION

Domain Feature extraction is one of the most complex tasks present in SA. It aims to deliver the identity of entities being referred. Then the classification task aims to express polarity of the statements that are determined to show the sentiment of statements.

The feature attributes of the subject influence subject quotient at different rate. It is termed as a feature quotient. For example the feature attributes of movie corresponds to,

Music Screenplay Graphics Acting Script, etc.

- Let us take one such feature into account and its feature quotient. The feature quotient varies for each attributes based on their functionality.
- Sample Feature-music:

Music is one of the features for movies; it makes in to account for a movie to provide the overall review. By extracting features for the subject, Sentiscore is provided in order to various aspects of the movie.

- Conversion to data frame:

Once the feature is extracted it is converted into data frames in order to arrange them in data tables, which is understood by the system.

- Conversion to corpus data:

The data frames are converted into corpus, which represents the extracted feature as a text. The text is then easy to analyze and for splitting the sentimental words through which the Sentiscore is calculated.

- Removal of stop words like "the" "is":

Corpus data contains lot of stop words like “the”, “is”, “was” etc. The above mentioned words are unnecessary for finding sentimental score of the feature extracted. By removing the stop words one can split the sentiment words from tweets.

- Conversion to character:

Corpus data for which the stop words are removed is converted as character in order to check polarity equivalent word in the positive or negative dictionary.

VI. SENTIMENTAL CLASSIFICATION

Classification of individual entity of the sentence to determine the overall sentiment expressed through that sentence.

Sentimental classification involves

- Determination of Positive/Negative words through identification of keywords.
- Assigning True/False value based on their polarity.
- Identification of Individual positive (mupos), negative (muneg) and overall score (musicscore).
- Calculation of overall sentimental score determined by the formula musicsentipos = (mupos/musicscore)*100. musicsentineg

= (muneg/musicscore)*100. pos<- readLines ("positive_words.txt")

The positive words are stored in the text file which is read into the variable pos by using the function readLines.

neg<- readLines ("negative_words.txt")

The positive words are stored in the text file which is read into the variable neg by using the function readLines.

- Converting into TRUE/FALSE:

The feature that is converted into character in feature extraction model is matched with words in the

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019

positive/negative word dictionary, if the match is found it is converted into TRUE/FALSE. Once the character is converted into the TRUE/FALSE, the sum of that TRUE/FALSE is calculated individually.

- Total score

The total score of the feature is calculated by adding individual sum of the polarity. This gives the value of total number of positive and negative sentiments that are present in the particular feature.

- Sentiment score:

In order to represent the sentiments in percentage form one needs to calculate it by using the formula musicsentipos = (mupos/musicscore)*100 (for music) which shows the overall sentimental score for that particular feature.

Result

Thus the process of sentimental analysis is carried out on twitter streaming data and the resulting values are represented as bar plot graph. Inference from the graph that is obtained here is tweets collected from real time streaming data about Movies in sentiment plot is high towards direction features. This signifies that direction features occupy a higher priority in movies. The Graph actually draws a conclusion on various issues which contribute in a movie's Success through Sentiment Analysis.

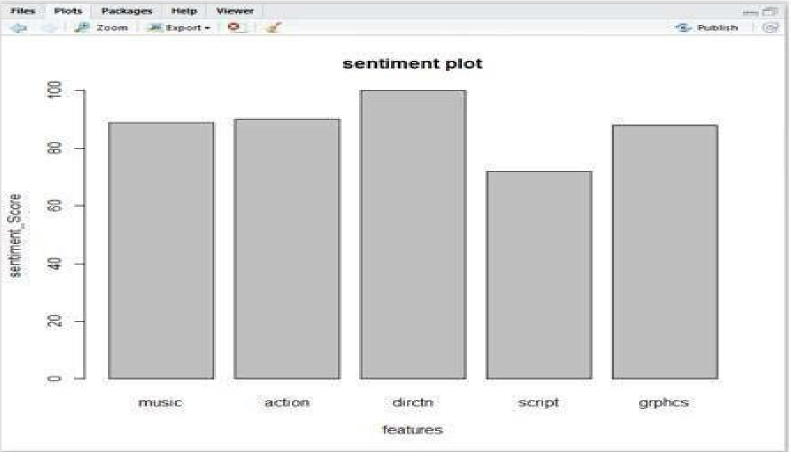


Fig. 1.5.1: Graph of Sentimental Analysis on Twitter

VII. Conclusion And Future Work

To conclude, the paper aims to deliver the analysis of sentimental perspective expressed through the tweets. To do this we have followed the following stages,

1. Tweet retrieval
2. Data pre-processing
3. Feature extraction
4. Sentimental classification and
5. Knowledge representation

In addition to the above mentioned, fuzzy logic movie recommender system based on sentiments is the proposed work which can classify movies based on the sentiments gathered. This can be the future work of the application.

REFERENCES

[1] AbinashTripathy, Ankit Agarwal, Santanu Kumar Rath (2015). Classification of Sentimental reviews using Machine Learning Techniques. *Procedia Computer Science,* Vol.57, pp.821- 829.

[2] Hassan Saif, Yulan He and Harith Alani. (2012).Semantics sentiment analysis of twitter, The Semantic Web – ISWC 2012,Vol.7469, Lecture Notes in Computer Science pp 508-524.

[3] Bo Pang and Lillian Lee (2004).A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Article No.* 271.

[4] Akish Kumar and Teeja Mary Sebastian (2012), Sentiment Analysis on Twitter. *IJCSI International Journal of Computer Science Issues,* Vol. 9, Issue 4, No 3.

[5] Efstratios Kontopoulos, Christos Berberdis, Theologos Dergiades (2013).Ontology-based Sentiment Analysis of Twitter Posts. *Expert Systems with applications,* Vol 40(10), pp.4065– 4074.

[6] Ruby Prabowo, Mike Thelwall. (2009). Sentiment Analysis: A combined approach. *Journal of Informetrics,* Vol. 3 (2) pp.143-157.

[7] Isidro penalver-Martinez, Francisco Garcia-Sanchez, Rafael Valencia- gracia, Miguel Angel Rodriguez-Garcia, Valentin Moreno, Anabel Fraga, Jose Luis Sanchez-Cervantes. (2014). Feature-based opinion mining through ontologies. *Expert Systems with Applications,* Vol. 41(13), pp.5995– 6008.

[8] Efthymioskouloumpis, Theresa Wilson, Johanna Moore. (2011) Twitter Sentiment Analysis: The good the Bad and the OMG, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.*

[9] Jantimaponpinij and Adhithya.k.ghose (2008).Ontology based classification methodology for online consumer reviews. *WI-IAT '08 Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* – Vol. 01, pp.518- 524.

[10] Lili Zhao and Chunoing Li (2014). Ontology based opinion mining for movie reviews, KSEM'09 *Proceedings of the 3rd International Conference on Knowledge Science, Engineering and Management,* pp.204-214.

[11] Walaamedhat, Ahamehassan, hodakorashy (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal,* Vol.5 (4), pp. 1093 – 1113.