*International Journal of Engineering Inventions*
*ISSN: 2278-7461, www.ijeijournal.com*
*Volume 1, Issue 4 (September 2012) PP: 06-09*

# Natural Language Processing For Content Analysis in Social Networking

**Mr. A. A. Sattikar[1], Dr. R. V. Kulkarni[2]**
[1] *V. P. Institute of Management Studies & Research, Sangli, Maharashtra, India*
[2] *Shahu Institute of Business Education & Research (SIBER), Kolhapur, Maharashtra, India*

***Abstract*—After going through review of almost twenty-five researches on security and privacy issues of social networking, researchers observed that users have strong expectations for privacy on Social Networking web sites as in Social Networking blogs or posts content is surrounded by a very high degree of abusive or defaming content .Social networking has emerged as the most important source of communication in the world. But a huge controversy has continued in full force over supervising offensive content on internet stages. Often the abusive content is interspersed with the main content leaving no clean boundaries between them. Therefore, it is essential to be able to identify abusive content of posts and automatically rate the content according the degree of abusive content in it. While most existing approaches rely on prior knowledge of website specific templates and hand-crafted rules specific to websites for extraction of relevant content, HTML DOM analysis and visual layout analysis approaches have sometimes been used, but for higher accuracy in content extraction, the analyzing software needs to mimic a human user and under- stand content in natural language similar to the way humans intuitively do in order to eliminate noisy content. In this paper, we describe a combination of HTML DOM analysis and Natural Language Processing (NLP) techniques for rating the blogs and posts with automated extractions of abusive contents from them.***

***Keywords — BLOG, HTML, NAIVE BYES, NLP, POST***

## I. INTRODUCTION

Natural Language Processing (NLP) is both a modern computational technology and a method of investigating and evaluating claims about human language itself. Some prefer the term Computational Linguistics in order to capture this latter function, but NLP is a term that links back into the history of Artificial Intelligence (AI), the general study of cognitive function by computational processes, normally with an emphasis on the role of knowledge representations, that is to say the need for representations of our knowledge of the world in order to understand human language with computers. Natural Language Processing (NLP) is the use of computers to process written and spoken language for some practical, useful, purpose: to translate languages, to get information from the web on text data banks so as to answer questions, to carry on conversations with machines, so as to get advice about, say, investments and so on. These are only examples of major types of NLP, and there is also a huge range of small but interesting applications, e.g. getting a computer to decide if one film story has been rewritten from another or not. NLP is not simply applications but the pure technical methods and processes that the major tasks mentioned above divide up into Machine Learning techniques which is automating the construction and adaptation of machine dictionaries, modelling human agents' beliefs and desires etc. The last task is closer to Artificial Intelligence, and is an important component of NLP. If computers are to involve in realistic conversations: they must, like human, have an internal model of them they converse with."NLP goes by many names — text analytics, data mining, computational linguistics — but the basic concept remains the same. NLP relates to computer systems that process human language in terms of its meaning. Apart from common word processor operations that treat text like a sheer sequence of symbols, NLP considers the hierarchical structure of language: many words make a phrase, many phrases make a sentence and, ultimately, sentences convey messages. By analyzing language for its meaning, NLP systems have many other useful roles, such as correcting grammar, converting speech to text and automatically translating text between languages. NLP can analyze language patterns for understanding text. One of the most convincing ways NLP offers valuable intelligence is by tracking sentiment — the nature of a written message (tweet, Facebook update, etc.) — and tag that text as positive, negative or neutral. Much can be gained from sentiment analysis. Companies can target unsatisfied customers or, find their competitors' unsatisfied customers, and generate leads. These examples can be called as "actionable insights" and can be directly implemented into marketing, advertising and sales activities. Every day, people discuss brands many of times across social media sites. Companies want a little of that to determine how their customer communicates, and more importantly, to discover important information that helps business. However, the sheer density of social conversations makes that difficult. Digital marketers and professionals are taking help of artificial intelligence tools like Natural Language Processing(NLP) to filter the social noise and properly connect with their target customers. If such tools really detect and analyse how a customer feels about a brand, is it possible with Natural Language Processing to analyse the blogs/posts of Social networking sites for detecting abused and defaming content? One of the long-term goals of artificial intelligence is to develop the programs that are capable of understanding and generating human language. A fundamental aspect of human intelligence seems to be not only the ability to use and understand natural language, but also its successful automation that have an incredible impact on

the usability and effectiveness of computers themselves. Understanding natural language is much more than parsing sentences into their individual parts of speech and looking those words up in a dictionary. Real understanding depends on broad background knowledge about the domain of conversation and the idioms used in that domain as well as an ability to apply general contextual knowledge to resolve the omissions and ambiguities that are a normal part of human speech.

A general text analysis process includes following sub tasks.

- Information retrieval or identification of a corpus is a preliminary step: collecting or identifying set textual materials, on the Web or held in a file system, database, or content management system, for analysis. Although some text analytics systems bound themselves to purely statistical methods, many others apply broader natural language processing, such as part of speech tagging, syntactic parsing, and other types of linguistic analysis.
- Named entity recognition is the use of statistical techniques to identify named text features: people, organizations, place names, stock ticker symbols, certain abbreviations, and so on. Disambiguation — the use of contextual clues — may be required to decide where, for instance, "Ford" refers to a former U.S. president, a vehicle manufacturer, a movie star (Glenn or Harrison), a river crossing, or some other entity.
- Recognition of Pattern Identified Entities: Features such as telephone numbers, e-mail addresses, and quantities (with units) can be discerned through regular expression or other pattern matches.
- Coreference: identification of noun phrases and other terms that refer to the same object. • Relationship, fact, and event Extraction: identification of associations among entities and other information in text
- Sentiment analysis involves discerning subjective material and extracting various forms of attitudinal information: sentiment, opinion, mood, and emotion. Text analytics techniques are helpful in analyzing sentiment at the entity, concept, or topic level and in distinguishing opinion holder and opinion object.
- Quantitative text analysis is a set of techniques stemming from the social sciences where either a human judge or a computer extracts semantic or grammatical relationships between words in order to find out the meaning or stylistic patterns of a casual personal text for the purpose of psychological profiling etc.

## II. ROLE OF NLP IN SOCIAL MEDIA

The rise of social media such as blogs and social networks has raised interest in sentiment analysis. With the increase in reviews, ratings, recommendations and other forms of online expression, online opinion has turned into a kind of virtual currency for businesses looking to sell their products, identify new opportunities and manage their reputations. As businesses look to automate the process of filtering out the noise, understanding the conversations, identifying the relevant content and actioning it properly, many are now looking to the field of sentiment analysis. If web 2.0 was all about democratizing publishing, then the next stage of the web may well be based on democratizing data mining of all the content that is getting published. One step towards this aim is accomplished in research, where several research teams in universities around the world currently focus on understanding the dynamics of sentiment in e-communities through sentiment analysis. The CyberEmotions project, for instance, recently identified the role of negative emotions in driving social networks discussions. Sentiment analysis could therefore help understand why certain e-communities die or weaken away (e.g., MySpace) while others seem to grow without limits (e.g., Facebook). However Social Networking Sites now days are using such techniques to get more smart. Facebook is using natural language processing to group posts in your News Feed, and link to a Page relevant to the topic that is being discussed. If more than one of our friends post about the same topic, and it has a Page on the social network, the posts will be grouped under a Posted About story, even if your friends don't explicitly tag the Page. The story is posted in the following format: "[Friend] and [x] other friends posted about [Page]" where the last part is a link to the Page in question.

But besides this tracking sentiment in the popular social networking sites and blogs has long been of need of today. With the availability of blogs and posts via social networking, it is now possible to automate some aspects of this process. A system can be developed that will use active machine learning technique to monitor sentiment in blogs and posts from popular social networking sites. The proposed system has two novel aspects. Firstly, it generates an aggregated posts feed containing diverse set of messages relevant to the concerned subject. This allows users to read the posts/blogs from their preferred social networking sites and annotate these posts/blogs as "positive" or "negative" through embedded links. Secondly, the results of this manual annotation process are used to train a supervised learner that labels a much larger set of blogs/posts. The annotation and classification trends can be subsequently tracked online. The main motivation for applying machine learning techniques in this context is to reduce the annotation effort required to train the system. Annotates can only be asked to annotate approximately ten blogs/posts per day where the remaining articles are classified using a supervised learning system trained on the smaller set of manually annotated articles. A number of machine learning techniques can be used in the system.

In recent years, blogs have become increasingly popular and have changed the style of communications on the Internet. Blogs allow readers to interact with bloggers by placing comments on specific blog posts. The commenting behavior not only implies the increasing popularity of a blog post, but also represents the interactions between an author and readers. To extract comments from blog posts is challenging. Each blog service provider has its own templates to present the information in comments. These templates do not have a general specification about what components must be provided in a comment or how many complete sub-blocks a comment is composed of. HTML documents are composed of various kinds of tags carrying structure and presentation information, and text contents enwrapped by tags. The input to pattern identification is an encoded string from an encoder. Each token in the string represents an HTML tag or a non-tag text. The algorithm scans the tokens. When encountering a token that is likely to be the head of a repetitive pattern (called a "rule" hereafter too), the subsequent tokens are examined if any rules can be formed. Many kinds of irrelevant comments are posted. For example, spam comments may carry advertisements with few links. Besides, commenter may just leave a message for greeting. Identifying relevant comments is an important and challenging issue for correctly fining the opinion of readers. Natural Language Processing (NLP) can be leveraged in any situation where text is involved. NLP involves a series

of steps that make text understandable (or computable). A critical step, lexical analysis is the process of converting a sequence of characters into a set of tokens. Subsequent steps leverage these tokens to perform entity extraction (people, places, things), concept identification and the annotation of documents with this and other information. From the blogs/posts, a query selection process selects a diverse set of blogs/posts for manual annotation. The remainder is classified as "positive" or "negative" by a Bayes classifier based on the comparison with manual annotation. Naïve Bayes classifier would be effective for producing aggregate sentiment statistics analysis with due care that should be taken in the training process. The Bayesian Classifier is capable of calculating the most probable output depending on the input. It is possible to add new raw data at runtime and have a better probabilistic classifier. A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Typical Naive Bayes algorithm has following different methods used in it.

```
use Algorithm::NaiveBayes;
my $nb = Algorithm::NaiveBayes->new;
$nb->add_instance
(attributes => {foo => 1, bar => 1, baz => 3},
label => 'sports');

$nb->add_instance
(attributes => {foo => 2, blurp => 1},
label => ['sports', 'finance']);
... repeat for several more instances, then:
$nb->train;

# Find results for unseen instances
my $result = $nb->predict
(attributes => {bar => 3, blurp => 2});
```

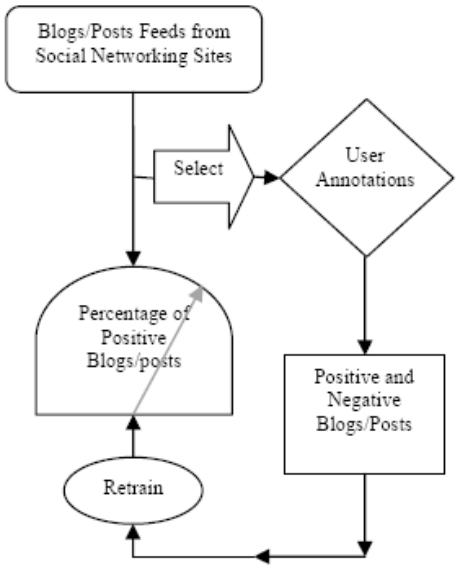


**Fig.1. Workflow of the Sentiment Analysis System**

The methods used in above algorithm can be explained as follows.

new()

Creates a new Algorithm::NaiveBayes object and returns it.
The following parameters are accepted:

add_instance( attributes => HASH, label => STRING|ARRAY )

Adds a training instance to the categorizer. The attributes parameter contains a hash reference whose keys are string attributes and whose values are the weights of those attributes. For instance, if you're categorizing text documents, the attributes might be the words of the document, and the weights might be the number of times each word occurs in the document.

The label parameter can contain a single string or an array of strings, with each string representing a label for this instance. The labels can be any arbitrary strings. To indicate that a document has no applicable labels, pass an empty array reference.

train()

Calculates the probabilities that will be necessary for categorization using the predict() method.

predict ( attributes => HASH )

Use this method to predict the label of an unknown instance. The attributes should be of the same format as you passed to add_instance(). predict() returns a hash reference whose keys are the names of labels, and whose values are the score for each label. Scores are between 0 and 1, where 0 means the label doesn't seem to apply to this instance, and 1 means it does.

As with most computer systems, NLP technology lacks human-level intelligence, at least for the likely future. On a text-by-text basis, the system's conclusions may be wrong — sometimes very wrong. For instance, the tweeted phrase "You're killing it!" may either mean "You're doing great!" or "You're a terrible gardener!" No automated sentiment analysis that currently exists can handle that level of nuance. Furthermore, certain expressions ("ima") or abbreviations ("#ff") fool the program, especially when people have 140 characters or less to express their opinions, or when they use slang, profanity, misspellings and neologisms.

Finally, much of social media interaction is personal, expressed between two people or among a group. Much of the language reads in first or second person ("I," "you" or "we"). This type of communication directly contrasts with news or brand posts, which are likely written with a more detached, omniscient tone. Furthermore, each of the above social media participants likely varies their language when they choose to post to Twitter vs. Facebook vs. Tumblr. Last but not least, the English language and intonation differs hugely based on the source and the forum.

## III. CONCLUSION

NLP is a tool that can help protect your privacy providing insight into the blogs and posts. However, it is not meant to replace human intuition. In social media environments, NLP helps to cut through noise and extract vast amounts of

data to help understand customer perception, and therefore, to determine the most strategic response. The challenges to producing useful applications of content analysis of Blogs/Posts, particularly within the context of automated analyses, are substantial. However, the benefits of a such analysis are even more substantial, including greater confidence in knowledge and the ability to predict future outcomes. In partial pursuit of such a manifesto, this paper mentions briefly algorithmic proposal for analysing contents of blogs and posts in social networking using natural language processing with supplemental user involvement.

## REFERENCES


1. Papadopoulos, E (2001), The relationship between the Internet financial message boards and the behavior of the stock market, ProQuest, http://sunzi.lib.hku.hk/ER/detail/hkul/3067352, retrieved 2011-09-23.
2. a b Peter Turney (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics. pp. 417–424. arXiv:cs.LG/0212032.
3. Bo Pang; Lillian Lee and Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86. http://www.cs.cornell.edu/home/llee/papers/sentiment.home.html.
4. Minqing Hu; Bing Liu (2004). "Mining and Summarizing Customer Reviews". Proceedings of KDD 2004.. http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.
5. Bo Pang; Lillian Lee (2004). "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts". Proceedings of the Association for Computational Linguistics (ACL). pp. 271–278. http://www.cs.cornell.edu/home/llee/papers/cutsent.home.html.
6. a b Bo Pang; Lillian Lee (2005). "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales". Proceedings of the Association for Computational Linguistics (ACL). pp. 115–124. http://www.cs.cornell.edu/home/llee/papers/pang-eestars. home.html.
7. Bing Liu; Minqing Hu and Junsheng Cheng (2005). "Opinion Observer: Analyzing and Comparing Opinions on the Web". Proceedings of WWW 2005.. http://www.cs.uic.edu/~liub/FBS/sentimentanalysis.html.
8. Kim, S.M. & Hovy, E.H. (2006). "Identifying and Analyzing Judgment Opinions.". Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLTNAACL 2006). New York, NY.. http://acl.ldc.upenn.edu/P/P06/P06-2063.pdf.
9. a b Benjamin Snyder; Regina Barzilay (2007). "Multiple Aspect Ranking using the Good Grief Algorithm". Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL). pp. 300–307.
10. Rada Mihalcea; Carmen Banea and Janyce Wiebe (2007). "Learning Multilingual Subjective Language via Cross-Lingual Projections". Proceedings of the Association for Computational Linguistics (ACL). pp. 976–983. http://www.cse.unt.edu/~rada/papers/mihalcea.acl07.pdf.
11. a b Lipika Dey , S K Mirajul Haque (2008). "Opinion Mining from Noisy Text Data". Proceedings of the second workshop on Analytics for noisy unstructured text data, p.83-90. http://portal.acm.org/citation.cfm?id=1390763&dl=GUIDE&coll=GUIDE&CFID=92244761&CFTOKEN=30578437.
12. Pang, Bo; Lee, Lillian (2008). "4.1.2 Subjectivity Detection and Opinion Identification". Opinion Mining and Sentiment Analysis. Now Publishers Inc. http://www.cs.cornell.edu/home/llee/opinion-miningsentiment- analysis-survey.html.
13. Fangzhong Su; Katja Markert (2008). "From Words to Senses: a Case Study in Subjectivity Recognition". Proceedings of Coling 2008, UK. http://www.comp.leeds.ac.uk/markert/Papers/Coling2008.pdf.
14. a b Wright, Alex. "Mining the Web for Feelings, Not Facts", New York Times, 2009-08-23. Retrieved on 2009-10-01.
15. Kirkpatrick, Marshall. ", ReadWriteWeb, 2009-04-15. Retrieved on 2009-10-01.
16. Bing Liu (2010). "Sentiment Analysis and Subjectivity". Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010. http://www.cs.uic.edu/~liub/FBS/sentimentanalysis.html.
17. Michelle de Haaff (2010), Sentiment Analysis, Hard But Worth It!, CustomerThink, http://www.customerthink.com/blog/sentiment_analysis_hard_but_worth_it, retrieved 2010-03-12.
18. Thelwall, Mike; Buckley, Kevan; Paltoglou, Georgios; Cai, Di; Kappas, Arvid (2010). "Sentiment strength detection in short informal text". Journal of the American Society for Information Science and Technology 61 (12): 2544–2558. http://www.scit.wlv.ac.uk/~cm1993/papers/SentiStrengthPreprint.doc.
19. A. A. Sattikar, Dr. R. V. Kulkarni (2011),"A review of Security and Privacy Issues in Social Networking", International Journal of Computer Science and Information Technologies, Vol. 2 (6) , 2011, 2784-2787
20. A. A. Sattikar, Dr. R. V. Kulkarni (2012),"A Role of Artificial intelligence in Security and Privacy Issues in Social Networking", International Journal of Computer Science and Engineering Technologies, Vol. 2 (1) , 2012, 792-795