Review paper on sentiment analysis on social networking sites

[1]Snehal P. Patil, [2]Prof.S.G.Tuppad

[1]Student, [2]Assistant Professor
Department of Computer Science and Engineering,
MSS's College of Engineering and Technology, Jalna,India

***Abstract*—This paper describes the approach we take to social media analysis, combining the exploration of the opinion of text and centered on the recognition of entities and events. We examine a particular use case, which is to help archivists select materials for inclusion in a social media archive to preserve community memories, moving towards structured preservation around semantic categories. The textual approach we adopt is rule based and relies on a number of sub-components, taking into account issues inherent in social media such as noisy non-grammatical text, use of insulting words, short language popularly called as SLANG, and so on. In order to resolve the ambiguity and provide additional contextual information, we propose two major innovations in this work: first, the novel combination of tools for extracting texts and multimedia opinions; and second, the adaptation of NLP tools for the analysis of opinion specific to the problems of social media.**

Index Terms— sentiment analysis, slangs,combinatorial and categorical grammar, tokenization.

I. INTRODUCTION

With the Internet explosion there is an abundance of data available online, they can be digital or text file and they can be structured, semi-structured or unstructured. Approaches and techniques for applying and extracting useful information [1] from these data have been the main focus of many researchers and practitioners in recent times. The advancement of computer technology as well as numerous techniques and tools of recovery have been proposed according to different types of data. In addition to the exploration of data and texts, there has been a growing interest in non-topical text analysis in recent years. The analysis of feeling is one of them. The analysis of feelings, also known as the exploration of opinion, consists of identifying and extracting subjective information in source materials that can be positive, neutral or negative [2]. Using appropriate mechanisms and techniques, this vast amount of data can be transformed into information to support operational, managerial and strategic decision-making [8]. The analysis of feelings aims at identifying and extracting opinions and attitudes from a given piece of text to a specific subject [11]. There has been a lot of progress on the analysis of conventional text feeling that is usually found in open forums, blogs and typical review channels. However, the analysis of the feeling of microblogs like Twitter is considered a much more difficult problem because of the unique characteristics of microblogs (for example, the short duration of status updates and language variations).

The widespread availability of the Internet has allowed people to express opinions that are much farther than ever it was possible. In addition, opinion data available on the Internet covers all recent trends, questions, opinions on each thinkable subject. These gigantic data [3] can be a potential source for evaluating a feeling. The analysis of feeling is the extraction of the feeling of all communication (verbal / non-verbal).

II. RELATED WORK

While much work has recently focused on analyzing social media to get an idea of what people think about current issues, there are still many challenges ahead. Advanced approaches to product reviews and so on are not necessarily appropriate to our task, partly because they generally operate in one narrow area and partly because the objective of public opinion is known In advance or at least to a limited subset (e.g. film titles, product names, companies, political parties, etc.). In general, sensing techniques can be roughly divided into lexicon based methods [11] and machine learning methods, [1]. Methods based on the lexicon are based on a lexicon of feeling, a collection of known and pre-compiled feeling terms.

Machine-learning approaches use syntactic and / or linguistic characteristics, and hybrid approaches are very common, sentiment lexicons playing a key role in most methods. For example, [15] establish the polarity of the examinations by identifying the polarity of the adjectives appearing therein, with a reported accuracy of about 10% greater than that of pure mechanical learning techniques. However, these relatively successful techniques often fail when they are moved to new areas or types of text, as they are inflexible with respect to the ambiguity of the terms of feeling. The context in which a term is used may change its meaning, especially for adjectives in the lexicons of feeling [9]. Several evaluations have shown the utility of contextual information [14] and have identified contextual words with a significant impact on the polarity of ambiguous terms [8].

Another bottleneck is the time creation of these sentiment dictionaries, although solutions have been proposed in the form of crowdsourcing techniques Their approach has much in common with a prior sentiment classifier constructed by [9], which also used unigrams, bigrams and POS tags, although the first one demonstrated by analysis that the distribution of some POS tags varies between positive and Negative posts. One reason for the relative scarcity of linguistic techniques for exploring social media opinion is probably due to difficulties in using NLP for low-quality text [2]; for example. The Stanford NER increases from

90.8% F1 to 45.88% when applied to a corpus of tweets [4]. There have been a number of recent work to detect sarcasm in tweets and other user-generated content [13, 11, 12, 3], with typical accuracy around 70-80%. These are mostly formed on a set of tweets with the hashtag #sarcasm and / or #irony, but all try to classify whether a sentence or tweet is sarcastic or not (and sometimes in a set of predefined sarcasms).

However, none of these approaches goes beyond the initial classification stage and therefore cannot predict how sarcasm will affect the expressed feeling. This is one of the issues we address in our work. Extracting the feeling of images is still an area of research that is in its infancy and not yet prolifically published. However, those published often use small sets of data for their truth on which to build SVM classifiers. Evaluations show that systems often respond a little better than chance to emotions driven from general images [7]. The implication is that the selection of characteristics for such a classification is difficult. [5] used a set of color characteristics to classify their small set of truth data to the ground, also using SVM, and publish an accuracy of about 87%. In our work, we develop this color-based approach to use other features and also use the wisdom of the crowd to select a large set of ground truth data. Other documents have begun to allude to the multimodal nature of image sentiment on the Web. Earlier works, such as [11], concern annotations of similar multimodal images, but not specifically for sentiment. They use latent semantic spaces to correlate the characteristics of the image and the text in a single space of characteristics. In this article we describe the work we have undertaken using text and images together to form sentiment for social media.

The analysis of feeling is in itself a major field of study under machine learning. The ideology used in this project is based on the underlying principles developed in [5] where the tweets were classified using unigram vectors and the training was carried out by remote supervision. The research in [5] elucidates that the use of emoticons as labels is effective in reducing dependencies in machine learning. The analysis of [5] is also based on a query term and characteristic reduction using algorithms such as Naive Bayes, Maximum Entropy and Support Vector Machines. The research work and further analysis were majorly conducted by Pang and Lee [3] this work was used to analyze the performance of various automated learning techniques in the field of cinema examination. It has also found implementations [11] as a subcomponent technology increasing with other systems like e-mails and online advertisements. Through enhanced natural language processing capabilities and tools, this domain is gaining in importance and improving application in various other areas.

Text mining data generally known as text mining is the process of gaining useful information from interesting and non-trivial models that are enormously available on Internet social networking sites or public forums. As we have seen [2], the techniques used in the extraction of text are inherited from the retrieval of information, extraction of information and natural language processing areas. It is also associated with algorithms and methods of knowledge discovery from databases (KDD) [3], data mining [4] and automatic learning techniques [5].

As noted in [6], textual information is classified into two categories: facts and opinions. Facts are only objective expressions about entities whereas opinions are subjective expressions that describe people's feelings toward entities. The analysis of feelings that is interchangeably also known as mine of opinion extracts feelings such as positive, negative and neutral from the written text. The authors of [7] used Naïve Bayes; classify maximum entropy and SVM to classify the reviews of films as positive or negative. Different combinations of features such as unigrams, unigrams and bigrams, unigram and POS (Part-Of Speech) labels have been used. The SVM technique surpassed the results of all other techniques. SVM obtained an accuracy of 81.6% in the unigram presence functionality. Automatic learning approaches in the field of feel analysis, such as the regression model as discussed in [8], are used to predict the usefulness of a revision. A semi-supervised method of performing a binary classification of text as positive or negative is described in [9].

III. PROPOSED WORK

It is much more natural for subjects to communicate their thoughts in the language of natural communication. While it is difficult to pursue people to fill out structured questionnaire, natural communication is easily accessible via various posts on social networks and microblogging sites. This approach works on the natural language data collected through this strategy.

With the natural language data collected from the previous step, the next step is to focus on reducing the semantic expressions by using the predefined semantic rules. For this purpose, a reduction function must be defined. This function repeatedly rewrites the semantic expressions based on the predefined rules. This reduction function can be specified using a functional language. Depending on pattern matching concepts, each predefined rule can be mapped one by one by declaring the function that accepts only the model of that rule. To reduce semantic expressions more effectively, certain additional functions must also be declared for structures that cannot be reduced. An additional identity function must be declared to process patterns that cannot be reduced by any other declared function.

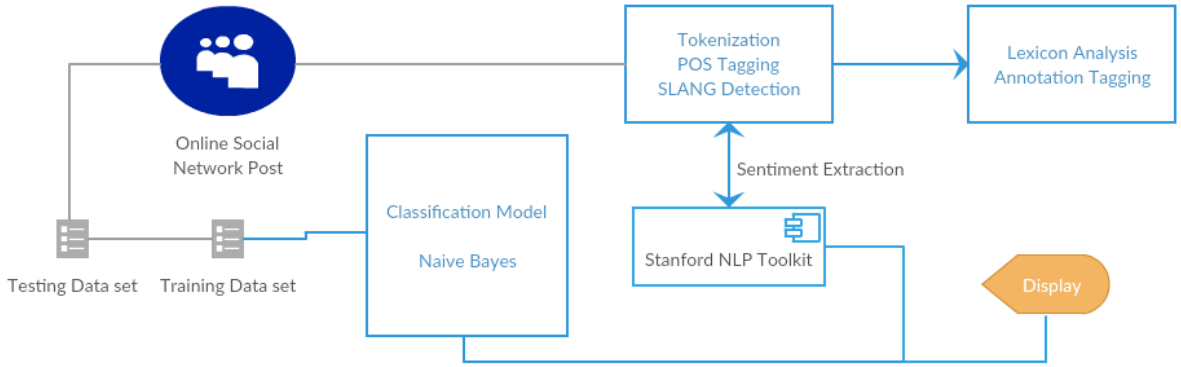

Figure 1. Proposed Architecture

IV. TOKENIZATION

Tokenization is the process of breaking a stream of text into sentences, words, symbols, or other meaningful elements called tokens. The objective of tokenization is the exploration of words in a sentence. Text data is only a textual interpretation or a block of characters at the beginning. In extracting information require the words from the dataset. We therefore need an analyzer that deals with the tokenization of documents. This can be trivial because the text is already stored in machine-readable formats. But there are still some problems that have been left, for example, the elimination of punctuation marks and other characters such as hooks, dashes, and so on. The main use of tokenization is the identification of significant keywords. Another problem are abbreviations and acronyms that need to be transformed into a standard form.

V. COMBINATORIAL AND CATEGORICAL GRAMMAR

A combinational categorical grammar maps from a lexical unit to a set of 2-tuples each are containing a lexical category and a semantic expression. The first tuple consists of lexical and phrasal categories and the second tuple is a set of semantic expressions. The set of lexical and phrasal categories follows an advanced structure as mentioned in [3] to incorporate modalities. A category is either primitive or compound. The set of primitive categories depends on the language and, for the English language; it consists of S (sentence), NP (syntagme), N (nom) and PP (prepositional sentence). The compound categories are defined recursively by the infix operators. This allows the formation of all other necessary lexical and phrasal categories. The operators remain associative, but to avoid confusion, internal composite categories are always encapsulated in parentheses. Syntactic categories form a type of system for semantic expressions, with a set of primitive types.

Combiners are only rules of inference of the proof system, since they take pairs of simple or multiple functions in the form of lexical category and semantic expression instances to produce new instances from the same set. The power of expression of the grammar depends on the combinators. The essential combinators mentioned in [14] constitute a context sensitive class grammar. The development of these combinatorial rules is presented in [15], with some modifications in the coordination conjunctions due to the modalities on infix operators. [14] and [15] also focuses on some additional combinators that identify few unique linguistic phenomena. Since the rules of inference are independent of language, some of the additional phenomena covered by the author in [14], [15] are either rare or do not exist.

VI. LEXICON ACQUISITION AND ANNOTATION

The existence of a lexicon with wide coverage is a must to implement the proposed method on the actual data. Considerable efforts have already been made to construct a combinational lexicon of categorical grammar from a corpus marked by a part of the speech (a corpus marked by POS). Each token is marked with its part of the word like name, adjective, verb, etc. in a corpus labeled POS. But this approach has major problems because it lacks appropriate structure as a bank of trees. Some existing lexicons such as CCGbank, compiled by [6] are based on techniques covered in [12]. This is nothing more than the translation of the entire Penn Treebank [8], containing more than 4.5 million chips, and where each sentence structure has been analyzed in full and annotated. The resulting lexicon has a high coverage where some entries are attributed to more than 100 different lexical categories. The authors of [6] calculated that the expected number of lexical categories per token is 19.2 for CCGBank. This implies that even searching for a short sentence (seven chips) should consider about more than 960 million possible marking. Therefore, this is not a feasible approach, although syntactic analysis can explore all possible inferences in polynomial time of the number of possible markings.

A possible solution should target the context in which the token appears to reduce marking. A machine learning approach based on a maximum entropy model presented in [9] estimates the probability that a token's will be assigned to a particular category, taking into account the characteristics of the local context. This is used to select a subset of possible categories for a token, by selecting categories with a probability in a factor of the highest probability category. [10] Show that the average number of lexical categories per token can be reduced to 3.8 while the analyzer still recognizes 98.4% of the invisible data. The authors of [30] present a complete analyzer that uses this method of marking and a series of methods that are log linear to accelerate the actual deduction once the management model has assigned a set of categories to each token. CCGBank, as well as toolchain is used as it can be freely used in education or research area.

The following is a brief review of some annotations of particular cases:

- Determinants: The significance of the determinants is comparatively less important when analyzing sentiment because it does not change the overall polarity of opinion with respect to any entity.
- Name: Generally, names must be managed by the generic algorithm. But there may be cases of multi-word names, where the partial name can be annotated by a list structure, which eventually captures the integer name.
- Verbs: Verbs in general are managed by the generic algorithm. But the particular case of binding verbs that are used to relate the subject to one or more predicative adjectives may be annotated with the identity function.
- Adjectives: They can be classified into different types according to their use in the sentence. The annotation is made with the change of the argument based on the lemma of the adjective which implies the implicit type conversion.
- Adverbs: Adverbs can be annotated mainly in the same way as adjectives. But here intensifiers and qualifiers that is adverbs that respectively increases or weakens the meaning must be scaled, based on the lemma.
- Relative prepositions and pronouns: They play a role in the argument of impact of partial sentences such as preposition sentences and relative clauses. These models should adhere to the whole sentence or clause.
- Conjunctions: They are annotated by an algorithm very similar to generic algorithms, which give a list structure instead of arguments function. The advantage of this annotation is that it allows any modification to bind on each of the conjugate sub phrases.

VII. Slang detection

People use Internet slang words such as "OMG" and "LOL" to express their feelings. The identification of slang feeling words can be an extraordinary benefit to accurately discover the hidden feeling in tweets and customer reviews.

Slang words (phrases) are those that are not present in dictionaries, while they are widely used to express feelings. Existing sentiment lexicons focus mainly on formal words, which do not contain an extended list of slang words. UrbanDictionary has an extensive list of slang words, while sentiment polarity is not available.

New slang words are created continuously. Therefore, it is essential to continue to extend the vocabulary of Slang Dictionary. Extension of Slang Dictionary consists of two steps: obtaining new slang words and labeling their polarity feeling; both steps can be automated.

VIII. Naïve Bayes classifier

A naive bayes classifier [15] is a simple probabilistic model based on the Bayes rule with a strong hypothesis of independence. The Naïve Bayes model implies a simplified conditional independence hypothesis. This is given a class (positive or negative), the words are conditionally independent of each other. This assumption does not significantly affect the accuracy of the text classification, but makes the classification algorithms very fast applicable to the problem. In our case, the probability of maximum likelihood of a word belonging to a given class is given by the expression:

$$P(x|c) = \frac{count of x in tweets of class c}{total number of words in class c} \quad (1)$$

Here, the xi is the individual words of the post tweet. The classifier delivers the class with the maximum a posteriori probability. We also remove duplicate words from tweets, they do not add any additional information; this type of naive bayes algorithm is called Bernoulli Naïve Bayes. The inclusion of the presence of a word instead of the count has been found to improve performance marginally, when there are a large number of training examples.

IX. Key Index Parameters for Result Classification

In information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, whereas recall (also called sensitivity) is the fraction of the relevant instances which are recovered. Accuracy and recall are therefore based on an understanding and measurement of relevance.

In simple terms, high accuracy means that an algorithm returned results significantly more relevant than irrelevant, while high recall means that an algorithm returned most relevant results.

The most important category measurements for binary categories are:

Accuracy:

$$P=TP/ (TP+FP) \quad (2)$$

Recall:

$$R=TP/ (TP+FN) \quad (3)$$

For classification purposes, the test data is pre-processed and a test data characteristic vector is formed. These test data are then introduced into the Bayes Naive algorithm with the training data to calculate the probability using the conditional Bayes Naive probability formula to obtain the polarity of the highest probability.

X. Conclusion

The analysis of feeling in the short and informal text is a fundamental problem for various fields. Although methods are proposed to solve this problem, an important challenge of identifying feeling in the informal / short text is the lack of lexical resources to understand the strength of the feeling of the slang words. To this end, we propose a web-based learning approach to build the first slang word dictionary, using available online resources. It is demonstrated that the Slang dictionary can actually improve the state-of-the-art informal text sense analysis tool, and it can be easily incorporated as an additional feeling lexicon. Future work includes the addition of new slang words to expand the Slang dictionary coverage and classification using other classification techniques to improve key performance indicators.

XI. Acknowledgment

I would like to thank my Project guide, Prof. S.G.Tuppad for her patient guidance, encouragement and advice she has provided throughout my time as her student and who responded to my questions and queries so promptly. I would also like to thank the Head of my Department Prof.G.P.Chakote for his constant support and guidance.I must express my gratitude to ourPrincipal Dr.S.K.Biradar, for his continued support and encouragement. I would like to thank my department and Faculty for providing the knowledge which allowed me to undertake this review.Finally, I am grateful to all the people who have supported me to complete the research work directly or indirectly.

References

[1] Bing Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, 2007.
[2] Apoorv Agarwal, BoyiXie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media, pages 30–38. Association for Computational Linguistics, 2011.
[3] Lillian Lee Bo Pang and ShivakumarVaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceeding EMNLP '02 Proceedings of the ACL- 02 conference on Empirical methods in natural language processing - Volume 10 Pages 79-86, pages 81–82, 2008.
[4] CNN Doug Gross. Survey: More americans get news from internet than newspapers or radio, 2010. [Online; accessed 9- July-2014].
[5] Alec Go, RichaBhayani, and Lei Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pages 1–12, 2009.
[6] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177. ACM, 2004.
[7] Leveragenewagemedia. Social media comparison infographic, 2013.[Online; accessed 9-July-2014].
[8] Bing Liu. Sentiment analysis and subjectivity.Handbook of natural language processing, 2:568, 2010.
[9] Laurent Luce. Twitter sentiment analysis using python and nltk, 2014. [Online; accessed 9-July-2014].
[10] Christopher D Manning and HinrichSch¨utze. Foundations of statistical natural language processing. MIT press, 1999.
[11] Bo Pang and Lillian Lee.Opinion mining and sentiment analysis. Foundations and trends in Information retrieval, 2(1- 2):1–135, 2008.
[12] Niek Sanders. Twitter sentiment corpus, 2014. [Online; accessed 9-July-2014].
[13] R. Feldman and J. Sanger, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data," First edition, Cambridge University Press, New York, 1995.
[14] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining," MIT Press,Cambridge, 1996.
[15] D. Hand, H. Mannila, and P. Smyth, "Principles of Data Mining," MIT Press, Cambridge, Massachusetts, 2001.