OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

AN IMPROVED ALGORITHM FOR IDENTIFYING SPAM IN ONLINE SOCIAL NETWORKING

Prof.B.K.Patil[1] , Ms. Monal N. Gamey [2]

Asst Professor[1], P.G. Student[2],Department of Computer Science & Engineering, Everest College of Engineering & Technology, Aurangabad, Maharashtra, India

monal_2525@yahoo.co.in

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.**

Keywords:- *Security, k-NN classifier, cloud databases, encryption*

---∴∴∴---

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

Text mining data generally known as text mining is the process of gaining useful information from interesting and nontrivial 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.

III PROPOSED ARCHITECTURE

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 micro blogging 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.

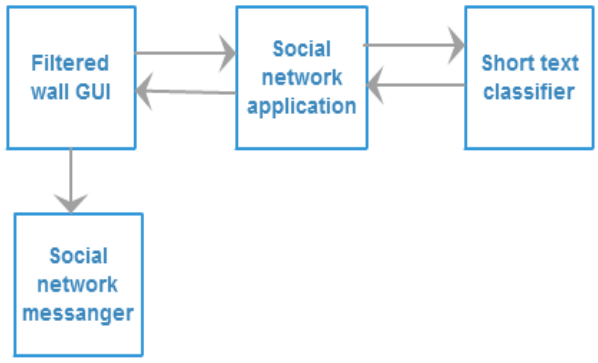

Figure 1 Work flow Process

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.

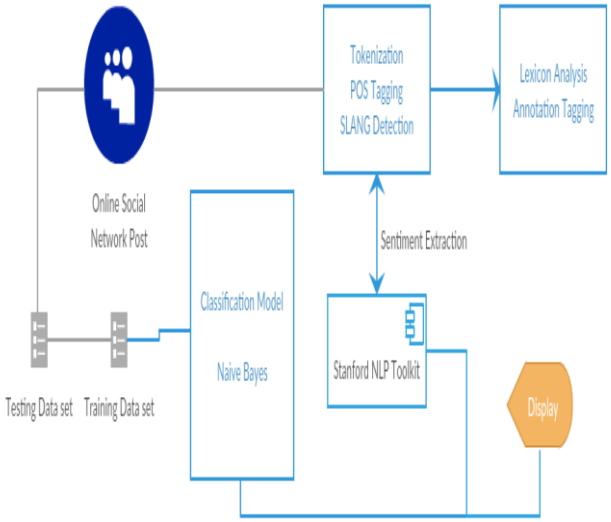

Figure 2 Proposed Architecture

Combinatorial and Categorical Grammar

A combinational categorical grammar maps from a lexical unit to a set of 2-tuples, each 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.

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.

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.

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. Urban Dictionary has an extensive list of slang words, while sentiment polarity is not available.

Naive 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:

$$P(x|c) = \frac{countofxintweetsofclassc}{totalnumberofwordsinclassc}$$

Here, the xi s are 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.

IV 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:

$$P = TP/(TP + FP)$$

Recall:

$$R = TP/(TP + FN)$$

Table 4.1: Results for the two stages of the proposed hierarchical classifier

Text Representation		Classification		
Features	**BoW TW**	**P**	**R**	**F_1**
Dp	-	37%	29%	33%
BoW	binary	69%	36%	48%
BoW	tf-idf	75%	38%	50%
BoW+Dp	binary	73%	38%	50%
BoW+Dp	tf-idf	74%	37%	49%
BoW+CF	binary	74%	58%	65%
BoW+CF	tf-idf	71%	54%	61%
BoW+CF+Dp	binary	74%	57%	64%
BoW+CF+Dp	tf-idf	76%	59%	66%

Table 4.2: Results of the proposed model in term of Precision (P), Recall (R) and F1

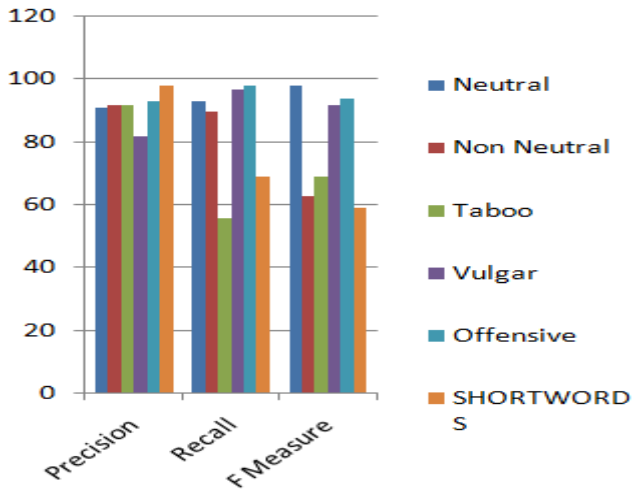

Figure 3 Graphical Result of Proposed System

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.

V 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.

REFERENCES

[1] P. Mell and T. Grance, "The nist definition of cloud computing (draft)," NIST special publication, vol. 800, p. 145, 2011.

[2] S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "Managing and accessing data in the cloud: Privacy risks and approaches," in CRiSIS, pp. 1 –9, 2012.

[3] P. Williams, R. Sion, and B. Carbunar, "Building castles out of mud: practical access pattern privacy and correctness on untrusted storage," in ACM CCS, pp. 139–148, 2008.

[4] P. Paillier, "Public key cryptosystems based on composite degree residuosity classes," in Eurocrypt, pp. 223–238, 1999.

[5] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted relational data." eprint arXiv:1403.5001, 2014.

[6] C. Gentry, "Fully homomorphic encryption using ideal lattices," in ACM STOC, pp. 169–178, 2009.

[7] C. Gentry and S. Halevi, "Implementing gentry's fullyhomomorphic encryption scheme," in EUROCRYPT, pp. 129– 148, Springer, 2011.

[8] A. Shamir, "How to share a secret," Commun. ACM, vol. 22, pp. 612–613, Nov. 1979.

[9] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy-preserving computations," in ESORICS, pp. 192–206, Springer, 2008.

[10] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in ACM Sigmod Record, vol. 29, pp. 439–450, ACM, 2000.
[11] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Advances in Cryptology (CRYPTO), pp. 36–54, Springer, 2000.
[12] P. Zhang, Y. Tong, S. Tang, and D. Yang, "Privacy preserving naive bayes classification," ADMA, pp. 744–752, 2005.
[13] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Information Systems, vol. 29, no. 4, pp. 343–364, 2004.
[14] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in IEEE ICDE, pp. 217–228, 2005.
[15] H. Hu, J. Xu, C. Ren, and B. Choi, "Processing private queries over untrusted data cloud through privacy homomorphism," in IEEE ICDE, pp. 601–612, 2011.
[16] M. Kantarcioglu and C. Clifton, "Privately computing a distributed k-nn classifier," in PKDD, pp. 279–290, 2004.
[17] L. Xiong, S. Chitti, and L. Liu, "K nearest neighbor classification across multiple private databases," in CIKM, pp. 840–841, ACM, 2006.
[18] Y. Qi and M. J. Atallah, "Efficient privacy-preserving k-nearest neighbor search," in IEEE ICDCS, pp. 311–319, 2008.
[19] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in ACM SIGMOD, pp. 563– 574, 2004.
[20] H. Hacig¨um¨ us¸, B. Iyer, C. Li, and S. Mehrotra, "Executing sql over encrypted data in the database-service-provider model," in ACM SIGMOD, pp. 216–227, 2002.
[21] B. Hore, S. Mehrotra, M. Canim, and M. Kantarcioglu, "Secure multidimensional range queries over outsourced data," The VLDB Journal, vol. 21, no. 3, pp. 333–358, 2012.
[22] W. K. Wong, D. W.-l. Cheung, B. Kao, and N. Mamoulis, "Secure knn computation on encrypted databases," in ACM SIGMOD, pp. 139–152, 2009.
[23] X. Xiao, F. Li, and B. Yao, "Secure nearest neighbor revisited," in IEEE ICDE, pp. 733–744, 2013.
[24] Y. Elmehdwi, B. K. Samanthula, and W. Jiang, "Secure k-nearest neighbor query over encrypted data in outsourced environments," in IEEE ICDE, pp. 664–675, 2014.
[25] M. Bohanec and B. Zupan. The UCI KDD Archive, 1997. http: //archive.ics.uci.edu/ml/datasets/Car+Evaluation.