ISSN:0975-9646

Sentiment Classification in Public Opinions using Multi-Domain Sentiment-Sensitive Thesaurus

Komal Paraswani , Professor P.P. Tribhuvan

Computer science and Engineering department,

Deogiri institute of engineering and management studies
Aurangabad ,Maharashtra, India

***Abstract*—The rise in social media use has changed the role of users from information receivers to information providers. As increasing numbers of people share their ideas, experiences, and opinions on the Web, sentiment analysis has become a popular topic for those who wish to understand public opinion from online data. Uptil now, research was only done on creating a model for the sentiments expressed by general public on the web and tracking them. We take a step ahead in identifying how sentiments vary over the web. We observed that emerging topics (named *foreground topics*) within a period that we consider for seeing variation in sentiments can be tracked down to identify the genuine reasons behind the variations, due to the relation between them. Based on this observation, we propose a Latent Dirichlet Allocation (LDA) based model that makes use of these foreground topics and removes the background topics that are untouched since long. To further enhance the results, we select the opinions that are most closely related to the foreground topics and develop another generative LDA-based model which assigns rank to the opinions on the basis of how famous the opinion is. *Sarcasm* is a form of speech act in which the speakers convey their message in an implicit way. The inherently ambiguous nature of sarcasm sometimes makes it hard even for humans to decide whether an utterance is sarcastic or not. Recognition of sarcasm can benefit many sentiment analysis NLP applications, such as review summarization, dialogue systems and review ranking systems. We move another step further to identify sarcasm and irony at the sentence as well as context level. The focus is on identifying irony in sentences containing *positive* predicates since these sentences are more exposed to irony, making their true polarity harder to recognize. We show this by exploring certain clues in the comments, such as emoticons, "lol's" and "rofl's", heavy punctuation marks, etc.**

***Keywords*—Natural Language Processing (NLP), Sentiment Analysis, Sarcasm Detection, Latent Dirichlet Allocation, Corpus**

INTRODUCTION

Others' opinions can be crucial when it's time to make a decision or choose among multiple options. When those choices involve valuable resources (for example, spending time and money to buy products or services) people often rely on their peers' past experiences. Until recently, the main sources of information were friends and specialized magazine or websites. Now, the "social web" provides new tools to efficiently create and share ideas with everyone connected to the World Wide Web. Forums, blogs, social networks, and content-sharing services help people share useful information. This information is unstructured, however, and because it's produced for human consumption, it's not something that's "machine processable."

Capturing public opinion about social events, political movements, company strategies, marketing campaigns, and product preferences is garnering increasing interest from the scientific community (for the exciting open challenges), and from the business world (for the remarkable marketing fallouts and for possible financial market prediction).

The resulting emerging fields are *opinion mining* and *sentiment analysis*. Although commonly used interchangeably to denote the same field of study, opinion mining and sentiment analysis actually focus on polarity detection and emotion recognition, respectively. Because the identification of sentiment is often exploited for detecting polarity, however, the two fields are usually combined under the same umbrella or even used as synonyms. Both fields use data mining and natural language processing (NLP) techniques to discover, retrieve, and distill information and opinions from the World Wide Web's vast textual information.

Companies use sentiment analysis to develop marketing strategies by assessing and predicting public attitudes toward their brand. Research and development focuses on designing automatic tools that crawl online reviews and condense the information gathered. Numerous companies already provide tools that track public viewpoints on a large scale by offering graphical summarizations of trends and opinions in the blogosphere. Developing opinion-tracking systems is commercially important.

Also, several tools already exist to help companies extract and analyze information from blogs about largescale trends in customers' opinions about products; those tools include SenticNet (http://sentic.net), Luminoso (http://luminoso.com), Factiva (http:// dowjones.com/factiva), Attensity (http://attensity.com), and Converseon (http://converseon.com). Most existing tools and research, however, are limited to polarity evaluation or mood classification according to a limited set of emotions. Such methods mainly rely on parts of text in which people explicitly express emotional states, and therefore the tools can't capture a reviewer's implicitly expressed opinion or sentiment.

The ability to identifying sarcasm and irony has got a lot of attention recently. The task of irony identification is not just interesting. Many systems, especially those that deal with opinion mining and sentiment analysis, can improve their

performance given the correct identification of sarcastic utterances[1][2].

One of the major issues within the task of irony identification is the absence of an agreement among researchers (linguists, psychologists, computer scientists) on how one can formally define irony or sarcasm and their structure. On the contrary, many theories that try to explain the phenomenon of irony and sarcasm agree that it is impossible to come up with a formal definition of these phenomena. Moreover, there exists a belief that these terms are not static but undergo changes and that sarcasm even has regional variations. Thus, it is not possible to create a definition of irony or sarcasm for training annotators to identify ironic utterances following a set of formal criteria. However, despite the absence of a formal definition for the terms irony and sarcasm, often human subjects have a common understanding of what these terms mean and can reliably identify text utterances containing irony or sarcasm.

There exist systems that target the task of automatic sarcasm identification[4]. However, the focus of this research is on sarcasm detection rather than on corpus generation[5]. Also, these systems focus on identifying sarcasm at the sentence level, or via analyzing a specific phrase, or "exploring certain oral or gestural clues in user comments, such as emoticons, onomatopoeic expressions for laughter, heavy punctuation marks, quotation marks and positive interjections".

THE PROBLEM OF SENTIMENT ANALYSIS

Liu et al. (2009) defines a sentiment or opinion as a quintuple-
"$<o_j, f_{jk}, so_{ijkl}, h_i, t_l>$, where o_j is a target object, f_{jk} is a feature of the object o_j, so_{ijkl} is the sentiment value of the opinion of the opinion holder hi on feature f_{jk} of object o_j at time t_l, so_{ijkl} is +ve,-ve, or neutral, or a more granular rating, hi is an opinion holder, t_l is the time when the opinion is expressed."

The analysis of sentiments may be document based where the sentiment in the entire document is summarized as positive, negative or objective. It can be sentence based where individual sentences, bearing sentiments, in the text are classified. SA can be phrase based where the phrases in a sentence are classified according to polarity.

Sentiment Analysis identifies the phrases in a text that bears some sentiment. The author may speak about some objective facts or subjective opinions. It is necessary to distinguish between the two. SA finds the subject towards whom the sentiment is directed. A text may contain many entities but it is necessary to find the entity towards which the sentiment is directed. It identifies the polarity and degree of the sentiment. Sentiments are classified as objective (facts), positive (denotes a state of happiness, bliss or satisfaction on part of the writer) or negative (denotes a state of sorrow, dejection or disappointment on part of the writer). The sentiments can further be given a score based on their degree of positivity, negativity or objectivity.

CHALLENGES FOR SENTIMENT ANALYSIS

Sentiment Analysis approaches aim to extract positive and negative sentiment bearing words from a text and classify the text as positive, negative or else objective if it cannot find any sentiment bearing words. In this respect, it can be thought of as a text categorization task. In text classification there are many classes corresponding to different topics whereas in Sentiment Analysis we have only 3 broad classes. Thus it seems Sentiment Analysis is easier than text classification which is not quite the case. The general challenges can be summarized as:

Implicit Sentiment and Sarcasm
A sentence may say something and mean something that is totally opposite to what is being said.

Domain Dependency
There exsist certain words whose polarity changes from domain to domain.

Thwarted Expectations
There can be situations where the writer writes a long paragraph is a single polarity. But the crucial last sentence of the paragraph totally reverts the meaning of the paragraph and so changes the polarity.

Pragmatics
Pragmatics such as capitalizations also play an important role in changing the polarity of the statement. A negative word can turn the sentence into positive using such pragmatics.

World Knowledge
Certain comments and statements might take reference of certain existing characters or places, etc to depict their sentiment. Our classifier need to be known to the knowledge of the world to identify such comparisons.

Subjectivity Detection
To actually identify sentiments, we need to in the first place identify whether the given statement is an opinion or not. Our classifier need not waste time in trying to classify statements that are mere facts and do not depict any sort of sentiment.

Entity Detection
A statement may consist of more than one entity. In such a situation it gets important to identify all the entities. This is so because it might happen that the statement turns out to be positive for one entity and negative for the other.

Negation
Negation is very difficult to identify in the sentiment analysis task. This is so because a statement can have a negative sentiment without the use of any negative sentiment words. Such statements prove to be a big challenge in NLP.

PROPOSED SYSTEM

CROSS-DOMAIN SENTIMENT CLASSIFICATION

The problem of classifying sentiments with cross-domain consideration has to face two major challenges. First is to train the classifier with the help of one or even multiple

domains that we call the source domains. Here it gets tough to identify which features from the source domains are related to which features in the target domains. Second challenge here is to apply the trained classifier on a domain that we haven't included in the source domains. Here we need a framework that learns using the information that shows how the features of source domain are related to the features of the target domain. It is very essential to look into these to challenges and overcome them to achieve a proper cross-domain sentiment classifier.

We model this problem of ours as one of feature expansion. Here we assign certain feature vectors to our working domain reviews. After that we assign additional related features to these feature vectors. This is done to reduce the mismatch that can occur between the source and target domains. Query expansion [9] in information retrieval [10], and document classification [11], and many such tasks make use of related features. But to the best of our knowledge, the cross-domain sentiment classification task has not been performed using the technique of feature expansion.

We begin by defining a domain which we denote as D. It represents a class that consists of entities in the world or a semantic concept. Then we consider two domains - the source domain (D_{src}) and the target domain (D_{tar}). Next, the (review, label) pairs (t, c) are chosen and added to the set of labeled instances, $L(D_{src})$, obviously from the source domain. Further we consider a set of unlabeled data from source domain $U(D_{src})$ and one from target domain $U(D_{tar})$. Finally after all the data is gathered, ie $L(D_{src})$, $U(D_{src})$, and $U(D_{tar})$, we begin the task of cross-domain sentiment classification by learning the classifier with the above data. Uptil now cross-domain classification was done using a single source domain. Here we consider multiple source domains that enhance the quality of our corpus and hence our classifier.

SARCASM IDENTIFICATION AND CLASSIFICATION

Our algorithm is semi-supervised. In the labelled input in the source domain we include labels for the level of sarcasm in the seed sentence. This labelling is done on a scale of 1 to 5, where 5 would denote a sentence that is totally sarcastic and 1 would be a sentence that is totally free of sarcasm. Next we extract a set of syntactic and pattern based features. And then we train our classifier discussed above to assign scores to the unlabeled examples.

CONCLUSION

Sentiment analysis, also called *opinion mining*, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. In today's world of internet where everybody is expressing their opinions on the web, the web proves to be an excellent source for mining opinions.

This report summarizes a method as to how to perform sentiment classification using a cross-domain sentiment sensitive dictionary. The methods used take into consideration various features. One of the most important feature considered is the implicit sarcasm in the text, i.e. parts of the text which show one opinion but imply a totally different opinion.

ACKNOWLEDGMENT

We hereby thank the authors listed in the References for the valuable information and survey statistics. We would like to thank anonymous reviewers for their valuable comments. We also thank you my guide and the participants from the user study for their support and early feedbacks on the design. We also sincerely thank the members of Internet Picture Dictionary group for allowing us to use their images. If I forget to mention the authors name or links which help me contribute their valuable information to me then I apologize to all of them.

REFERENCES

[1] Bo Pang and Lillian Lee," Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008), 1–135

[2] Paula Carvalho, Lu´ıs Sarmento, Mario J. Silva, and ´Eugenio de Oliveira. 2009. Clues for detecting irony ´in user-generated contents: Oh...!! It's "so easy" ;-). In *Proceeding of the 1st international CIKM workshop* on Topic-sentiment analysis for mass opinion, TSA '09, pages 53–56, New York, NY, USA. ACM.

[3] Roberto Gonzalez-Ibanez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of Association of Computational Linguistics.*

[4] Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM - A great catchy name: Semi-supervised recognition of sarcastic sentences in product reviews. In *Proceeding of AAAI Conference on Weblogs and Social Media (ICWSM–10)*, Washington, DC, USA, May.

[5] Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters.

[6] Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Targetdependent twitter sentiment classification. in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL- 2011). 2011.

[7] Ghani, Rayid, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. *Text mining for product attribute extraction.* ACM SIGKDD Explorations Newsletter, 2006. **8**(1): p. 41-48.

[8] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?:sentiment classification using machine learning techniques. In Proceedingsof Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). 2002.

[9] H. Fang, "A Re-Examination of Query Expansion using Lexical Resources," Proc. Ann. Meeting of the Assoc. Computational Linguistics (ACL '08), pp. 139-147, 2008.

[10] G. Salton and C. Buckley, Introduction to Modern Information Retreival. McGraw-Hill Book Company, 1983.

[11] D. Shen, J. Wu, B. Cao, J.-T. Sun, Q. Yang, Z. Chen, and Y. Li, "Exploiting Term Relationship to Boost Text Classification," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM '09), pp. 1637-1640, 20