Department of Information and Technology

Mini Project Report

ON

Twitter Sentiment Analysis

Submitted by:

Saksham Garg 1900910130107

AND

Sumit Julka 1900910130114

JSS Academy Of Technical Education , Noida

Dr. APJ Abdul Kalam Technical University , Lucknow , UP.

Session 2021-22

TABLE OF CONTENTS PAGE

ACKNOWLEDGEMENT

Presentation, inspiration and motivation have always played a key role in the success of any venture.

It gives us a great sense of pleasure to present this report on the topic of “**TWITTER SENTIMENT ANALYSIS – Supervised ML** for the project undertaken during B.Tech. third year. We owe a special debt of gratitude to our mentor for his constant support and guidance throughout the course of our work. His guidance has been very beneficial to us, especially whenever we were stuck and couldn’t make progress.

We simply couldn’t have done it without him. We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not least, we acknowledge all the resources that helped in completion of the report.

ABSTRACT

As we can see with the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too. The Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics, have discussions with different communities, or post messages across the world. There has been a lot of work in the field of sentiment analysis of twitter data. This survey focuses mainly on sentiment analysis of twitter data which is helpful to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. In this paper, we provide a survey and a comparative analysis of existing techniques for opinion mining like machine learning and lexicon-based approaches, together with evaluation metrics. Using various machine learning algorithms like Naive Bayes, Max Entropy, and Support Vector Machine, we provide research on twitter data streams.We have also discussed general challenges and applications of Sentiment Analysis on Twitter.

INTRODUCTION

Nowadays, the age of the Internet has changed the way people express their views, opinions. It is now mainly done through blog posts, online forums, product review websites, social media ,etc. Nowadays, millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinions and share views about their daily lives. Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides an opportunity for businesses by giving a platform to connect with their customers for advertising. People mostly depend upon user generated content over online to a great extent for decision making. For e.g. if someone wants to buy a product or wants to use any service, then they firstly look up its reviews online, discuss it on social media before making a decision. The amount of content generated by users is too vast for a normal user to analyze. So there is a need to automate this, various sentiment analysis techniques are widely used. Sentiment analysis (SA)tells users whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand about their products or services in such a way that it can be offered as per the user"s requirements. Textual Information retrieval techniques mainly focus on processing, searching or analyzing the factual data present. Facts have an objective component but there are some other textual contents which express subjective characteristics. These contents are mainly opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis (SA). It offers many challenging opportunities to develop new applications, mainly due to the huge growth of available information on online sources like blogs and social networks. For example, recommendations of items proposed by a recommendation system can be predicted by taking into account considerations such as positive or negative opinions about those items by making use of SA.

OBJECTIVE

After 10 years of documenting the world in 140 characters, Twitter now has more than 300m active users. It attracts a significant number of politicians, journalists, and celebrities.Twitter has obviously been used to raise awareness of political topics, spread political messages and coordinate collective action.But Twitter is also used to gauge public opinion, often producing a false sense of consensus or of how many people feel strongly about a topic (so-called Twitter storms). This is because users tend to connect with people who hold similar views to their own and are less likely to come across different issues and opinions.

Twitter has opened up a two-way communication between businesses and their customers. On the one hand this means it's easier for customers to complain to a company - and do so publicly. But it's also much quicker and easier for companies to reply and potentially resolve an issue, and can potentially even reduce customer support costs.

Newsrooms have long been dominated by the wires. Many journalists sit behind monitors, their eyes flicking towards the latest flashes in the corners of their screens. Twitter changed that, at least a bit. The flow of information around the world is no longer just controlled by the Associated Press or Reuters – it's being tweeted, too. Twitter's more than 300m users and every time a story breaks someone is there to post it, where it's shared almost instantly.

Twitter has changed celebrity culture beyond recognition. In the most basic sense, we are now able to follow the everyday life of a celebrity and, more importantly for many fans, communicate directly with them without strict control from their management. This means that a celebrity's image has become less about a set of fixed characteristics and more a shifting and organic performance in which the audience can participate.

Celebrities have to continuously maintain a persona that appears intimate, authentic and accessible. In some cases this really does mean instantly revealing their true thoughts in a way that wasn't previously possible. But for other celebrities this means crafting a product designed specifically for public consumption that is as tightly managed as a magazine photoshoot. This blurred boundary between image and reality has opened the way for a huge new kind of celebrity endorsement.

RELATED WORKS:

In recent years a lot of work has been done in the field of "Sentiment Analysis on Twitter " by a number of researchers. In its early stage it was intended for binary classification which assigns opinions or reviews to bipolar classes such as positive or negative only.

Pak and Paroubek(2010) proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, they developed a sentiment classifier based on the multinomial Naive Bayes method that uses features like N-gram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons.

Parikh and Movassate(2009) implemented two models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model.

Go and L.Huang (2009) proposed a solution for sentiment analysis for twitter data by using distant supervision, in which their training data consisted of tweets with emoticons which served as noisy labels. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. They concluded that SVM outperformed other models and that unigram was more effective as features.

Kamps used the lexical database WordNet to determine the emotional content of a word along different dimensions. They developed a distance metric on WordNet and determined the semantic polarity of adjectives.

Xia used an ensemble framework for Sentiment Classification which is obtained by combining various feature sets and classification techniques. In their work, they used two types of feature sets (Part-of-speech information and Word-relations) and three base classifiers (Naive Bayes, Maximum Entropy and Support Vector Machines) . They applied ensemble approaches like fixed combination, weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy.

Pre-processing and Feature Extraction

A tweet contains a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey work is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points, Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username) Correct the spellings; sequence of repeated characters is to be handled Replace all the emoticons with their sentiment. Remove all punctuations ,symbols, numbers Remove Stop Words Expand Acronyms(we can use a acronym dictionary) Remove Non-English Tweets.

The preprocessed dataset has many distinctive properties. In the feature extraction method, we extract the aspects from the processed dataset. Later this aspect is used to compute the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using models like unigram, bigram. Machine learning techniques require representing the key features of text or documents for processing. These key features are considered as feature vectors which are used for the classification task. Some example features that have been reported in literature are:

1. Words And Their Frequencies: Unigrams, bigrams and n-gram models with their frequency counts are considered as features. There has been more research on using word presence rather than frequencies to better describe this feature. Panget al. showed better results by using presence instead of frequencies.

2. Parts Of Speech Tags Parts of speech like adjectives, adverbs and some groups of verbs and nouns are good indicators of subjectivity and sentiment. We can generate syntactic dependency patterns by parsing or dependency trees.

3. Opinion Words And Phrases Apart from specific words, some phrases and idioms which convey sentiments can be used as features. e.g. cost someone an arm and leg.

4. Position Of Terms The position of a term within a text can affect how much the term makes difference in overall sentiment of the text.

5. Negation Negation is an important but difficult feature to interpret. The presence of a negation usually changes the polarity of the opinion.

Types of Learning

Machine learning based approach uses classification techniques to classify text into classes. There are mainly two types of machine learning techniques

3.1.1. Unsupervised learning:

It does not consist of a category and they do not provide the correct targets at all and therefore rely on clustering.

3.1.2. Supervised learning:

It is based on a labeled dataset and thus the labels are provided to the model during the process. These labeled dataset are trained to get meaningful outputs when encountered during decision- making. The success of both these learning methods mainly depends on the selection and extraction of the specific set of features used to detect sentiment. The machine learning approach applicable to sentiment analysis mainly belongs to supervised classification. In machine learning techniques, two sets of data are needed:

It is based on the usage of terms (seeds) that are usually collected and annotated manually. This set grows by searching the synonyms and antonyms of a dictionary. An example of that dictionary is WordNet, which is used to develop a thesaurus called SentiWordNet.

The corpus-based approach has the objective of providing dictionaries related to a specific domain. These dictionaries are generated from a set of seed opinion terms that grows through the search of related words by means of the use of either

statistical or semantic techniques.

Methods based on semantics such as the use of synonyms and antonyms or relationships from thesaurus like WordNet may also represent an interesting solution.

According to the performance measures like precision and recall, we provide a comparative study of existing techniques for opinion mining, including machine learning, lexicon-based approaches, cross domain and cross-lingual approaches, etc., as shown in Table 2.

Table 2. Performance Comparison Of Sentiment Analysis Methods

	Method	Data Set	Acc.	Author
Machine Learning	SVM	Movie reviews	86.40%	Pang, Lee[23]
	CoTraining SVM	Twitter	82.52%	Liu[14]
	Deep learning	Stanford Sentiment Treebank	80.70%	Richard[18]
Lexical based	Corpus	Product reviews	74.00%	Turkey
	Dictionary	Amazon's Mechanical Turk	---	Taboada[20]
Cross-lingual	Ensemble	Amazon	81.00%	Wan,X[16]
	Co-Train	Amazon, ITI68	81.30%	Wan,X.[16]
	EWGA	IMDb movie review	>90%	Abbasi,A.
	CLMM	MPQA,NTCIR,ISI	83.02%	Mengi
Cross-domain	Active Learning	Book, DVD, Electronics, Kitchen	80% (avg)	Li, S
	Thesaurus			Bollegala[22]
	SFA			Pan S J[15]

SCOPE AND FUTURE USAGE

1.Applications that use Reviews From Websites: The Internet has a large collection of reviews and feedback on almost everything. This includes product reviews, feedback on political issues, comments about services, etc. Thus there is a need for a sentiment analysis system that can extract sentiments about a particular product or services. It will help us to automate the provision of feedback or rating for the given product, item, etc. This would serve the needs of both the users and the vendors.

2. Applications as a Sub-component Technology: A sentiment predictor system can be helpful in recommender systems as well. The recommender system will not recommend items that receive a lot of negative feedback or fewer ratings. In online communication, we come across abusive language and other negative elements. These can be detected simply by identifying a highly negative sentiment and correspondingly taking action against it.

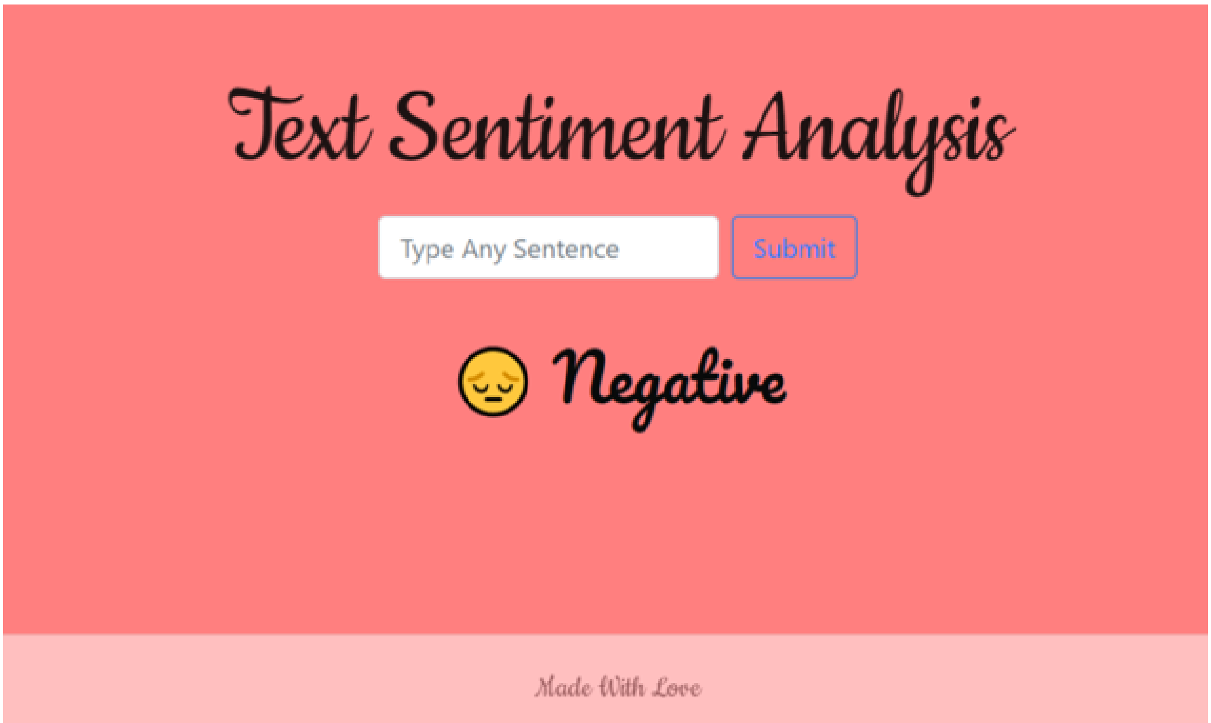
3. Applications in Business Intelligence It has been observed that people nowadays tend to look upon reviews of products which are available online before they buy them. And for many businesses, the online opinion decides the success or failure of their product. Thus, Sentiment Analysis plays an important role in businesses. Businesses also wish to extract sentiment from the online reviews in order to improve their products and in turn their reputation and help in customer satisfaction .

4. Applications across Domains: Recent Researches in sociology and other fields like medical, sports have also been benefited by Sentiment Analysis that show trends in human emotions especially on social media.

5. Applications In Smart Homes Smart homes are supposed to be the technology of the future. In future entire homes would be networked and people would be able to control any part of the home using a tablet device. Recently there has been a lot of research going on the Internet of Things(IoT). Sentiment Analysis would also find its way in IoT. For example, based on the current sentiment or emotion of the user, the home could alter its ambiance to create a soothing and peaceful environment. Sentiment Analysis can also be

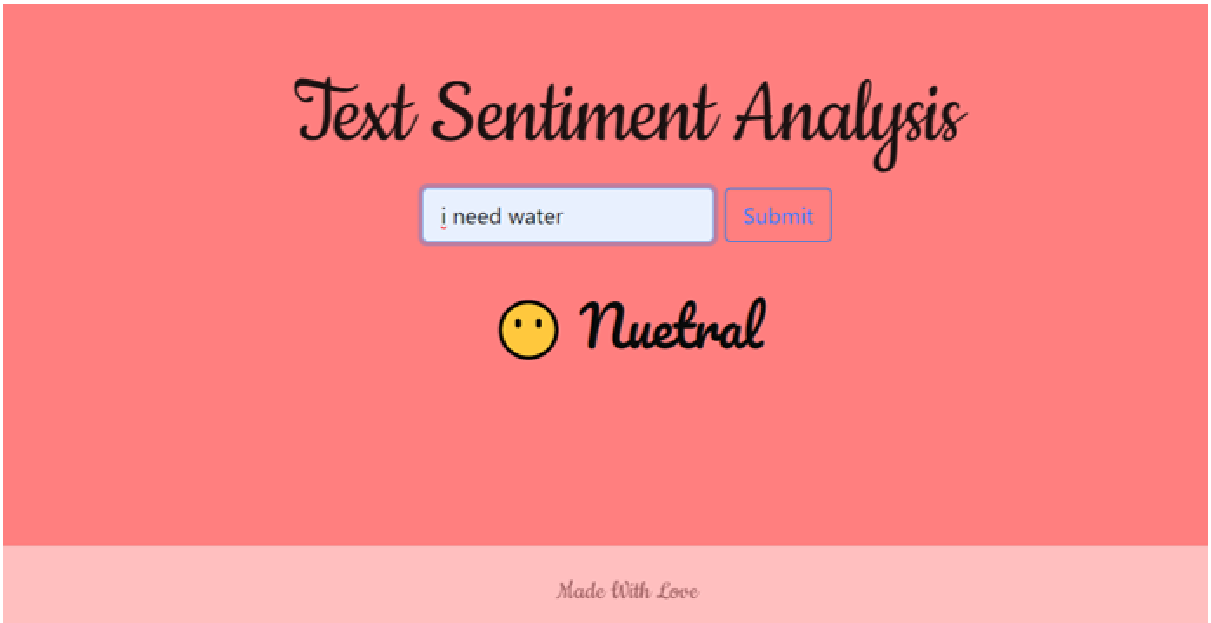
used in trend prediction. By tracking public views, important data regarding sales trends and customer satisfaction can be extracted.

Results and Snapshots



Text Sentiment Analysis

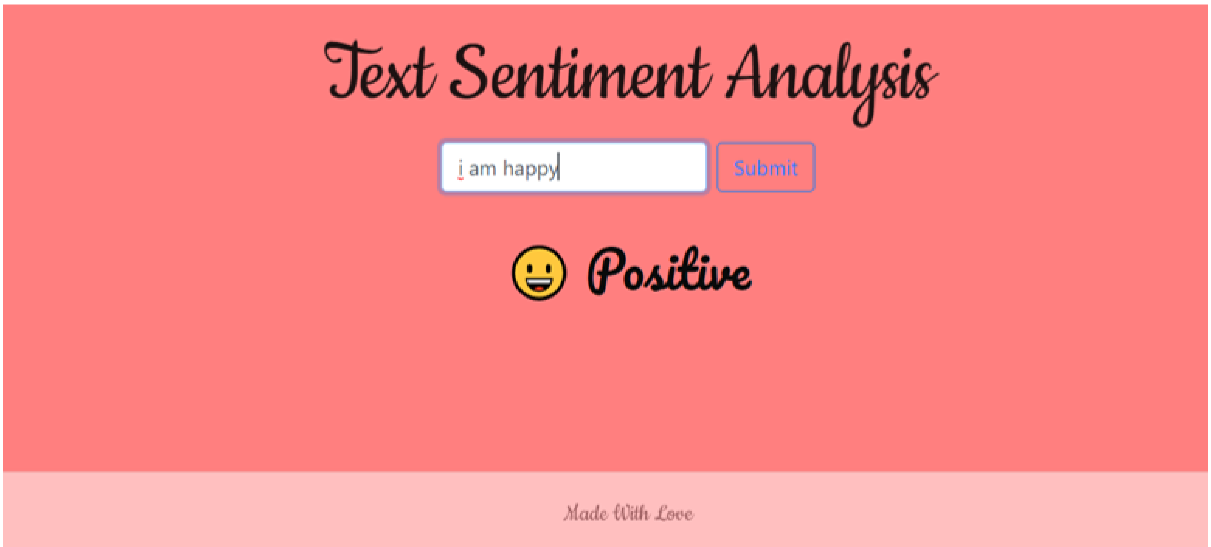
😔 Negative



Text Sentiment Analysis

i need water

😶 Nuetral



Text Sentiment Analysis

i am happy

😀 Positive

Made With Love

Conclusion

We have provided a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, together with cross domain and cross-lingual methods and some evaluation metrics. Research results show that machine learning methods, such as SVM and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very effective in some cases, which require few effort in human-labeled document .We also studied the effects of various features on classifier. We can conclude that with cleaner data, more accurate results can be obtained. Use of the bigram model provides better sentiment accuracy as compared to other models. We can focus on the study of combining machine learning methods into opinion lexicon methods in order to improve the accuracy of sentiment classification and adaptive capacity to a variety of domains and different languages.

References

[1] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[2] R. Parikh and M. Movassate, "Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi_cation Techniques",CS224N Final Report, 2009

[3] Go, R. Bhayani, L.Huang. "Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper,2009

[4] L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitter From Biased and Noisy Data". COLING 2010: Poster Volume,pp. 36-44.

[5] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[6] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011 , pp. 30-38

[7] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010

[8] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5,

[9] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

[10] Neethu M,S and Rajashree R," Sentiment Analysis in Twitter using Machine Learning Techniques" 4th ICCCNT 2013,at Tiruchengode, India. IEEE – 3166.

[11] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

[12] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," 2004.

[13] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[14] ZhunchenLuo, Miles Osborne, TingWang, An effective approachto tweets opinion retrieval", Springer Journal onWorldWideWeb,Dec 2013, DOI: 10.1007/s11280-013-0268-7.

[15] Liu, S., Li, F., Li, F., Cheng, X., &Shen, H.. Adaptive co-training SVM for sentiment classification on tweets. In Proceedings of the 22nd ACMinternational conference on Conference on information & knowledgemanagement (pp. 2079-2088). ACM,2013.

[16] Pan S J, Ni X, Sun J T, et al. “Cross-domain sentiment classification viaspectral feature alignment”. Proceedings of the 19th internationalconference on World wide web. ACM, 2010: 751-760.

[17] Wan, X..“A Comparative Study of Cross-Lingual SentimentClassification”. In Proceedings of the The 2012 IEEE/WIC/ACMInternational Joint Conferences on Web Intelligence and IntelligentAgent Technology-Volume 01 (pp. 24-31).IEEE Computer Society.2012

[18] Socher, Richard, et al. "Recursive deep models for semanticcompositionality over a sentiment Treebank." Proceedings of theConference on Empirical Methods in Natural Language Processing(EMNLP). 2013.

[19] Meng, Xinfan, et al. "Cross-lingual mixture model for sentimentclassification." Proceedings of the 50th Annual Meeting of theAssociation for Computational Linguistics Volume 1,2012

[20] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M..“Lexicon basedmethods for sentiment analysis”. Computational linguistics, 2011:37(2), 267-307.

[21] Li, S., Xue, Y., Wang, Z., & Zhou, G..“Active learning for cross-domainsentiment classification”. In Proceedings of the Twenty-Thirdinternational joint conference on Artificial Intelligence (pp. 2127-2133).AAAI Press,2013

[22] Bollegala, D., Weir, D., & Carroll, J.. Cross-Domain SentimentClassification using a Sentiment Sensitive Thesaurus. Knowledge andData Engineering, IEEE Transactions on, 25(8), 1719-1731,2013

[23] Pang, B.and Lee, L. “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”. 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04). 2004, 271-278.

[24] V. M. K. Peddinti and P. Chintalapoodi, “Domain adaptation in sentiment analysis of twitter,” in Analyzing Microtext Workshop, AAAI, 2011.