Sentiment Based Twitter Spam Detection

Nasira Perveen
Dept. of Comp. Science
Bahauddin Zakaria Univ. Multan
Pakistan

Malik M. Saad Missen
Dept. of Comp. Science and IT
The Islamia University of Bahawalpur
Pakistan

Qaisar Rasool
Dept. of Comp. Science
Bahauddin Zakaria Univ. Multan
Pakistan

Nadeem Akhtar
Dept. of Comp. Science and IT
The Islamia University of Bahawalpur
Pakistan

***Abstract*—Spams are becoming a serious threat for the users of online social networks especially for the ones like of twitter. twitter's structural features make it more volatile to spam attacks. In this paper, we propose a spam detection approach for twitter based on sentimental features. We perform our experiments on a data collection of 29K tweets with 1K tweets for 29 trending topics of 2012 on twitter. We evaluate the usefulness of our approach by using five classifiers i.e. BayesNet, Naive Bayes, Random Forest, Support Vector Machine (SVM) and J48. Naive Bayes, Random Forest, J48 and SVM spam detections performance improved with our all proposed features combination. The results demonstrate that proposed features provide better classification accuracy when combined with content and user-oriented features.**

***Keywords*—sentiment analysis; spam detection; twitter**

I. Introduction

Spam is a real threat to usefulness of the web. Spammers mask their content as useful or relevant content and hence is delivered to the user. The legitimate users consume this spam data considering it relevant to their information needs. Clay Shirky [2] remarked that a communication channel isnt worth its salt until the spammers descend.

Spams are not easy to stop. For several years, email services like Gmail, Microsoft and others have been successfully detecting spam emails but still spam emails are in circle on the web. These services have been reporting that email spamming has been up to 90 to 95 percent of the total email exchanges [3], [4], [5]. Even after successful detection of spams, companies are unable to stop spammers which ensures about the economical benefits spammers get when they trap a user clicking on a spam link. The severity of the threat posed by spamming has increased with the emergence of online social networks and twitter is one of the most popular online social network which has been highly affected by spam. twitter spamming is more threatening because its more targeted towards the trending topics of the twitter and hence bit easier to get penetrated especially because of hash-tag operator. Another fact that makes twitter a rather easier and fruitful target for spammers is its variety of audience. twitter users span across all sectors of life i.e. it can be the teachers or students, celebrities or politicians, marketers or customers or even general public. They belong to all age groups but most widely age group that uses twitter is between 55 to 64 years. There are about 60% users that access twitter from their cell phones [1]. twitter has 288 million monthly active members that make it widely growing social networking site. There are around 400 million tweets posted on daily bases, the average posts on twitter is 208 tweets per users account.

Due to this continuous distribution of information, a user faces many problems with search results that shares recurring and irrelevant information. This also can be very worrying at the times since a user has to scroll through the all information in direction to get an overall view of topic. Spam detection on the twitter network is difficult due to the noticeable usage of URLs, abbreviations, informal language and modern language concepts [6]. Old-style methods of detecting spam information fall short here. To date, study has been available on many techniques for detecting spams on twitter and blogs by using different features. After knowing the existing importance of spams on twitter, we take inspiration or motivation from this user need and decided to design and develop improved techniques to detect spams on twitter.

In this paper, we propose a spam detection approach for detecting spam tweets. This approach is based on sentimental features of a tweet. The idea is to exploit the philosophy that spammer use to force a user to click on a particular link. They definitly seek help of some motivational words (like 'the best web site', 'excellent service', etc) to make people believe in a certain tweet (examples of some spam tweets given in the table I. Results show that this exploitation of sentimental features proves fruitful.

TABLE I: Some Spam Tweet Examples

you'll laugh when you see this pic of you... tinyurl.coX/blah
you look like you lost weight in this video.. t.cX/blah
Was this blog you posted really necessary? tniX.biz/ad08 some kind of joke?
viagra,cialis,soma,tramadol and more. no prescription. ti.co/blah
Gain over 1,000 followers a week by using: ti.co/blah
wow this really works! i found out who stalks me :P go to 0rX.com/blah

The rest part of the paper is organized as the following

[1] http://blog.digitalinsights.in/social-media-facts-and-statistics-2013/0560387.html twitters Facts 2013

sections. In section II, we highlight some of the previous works done while in section III, we discuss proposed features. In Section IV-A, we describe the data collection used for experimentation. In Section IV, we describe our experimental results and comparisons of different features combinations and the conclusion is described in Section V.

II. Related Work

In this section, we describe several work related to spam detection on twitter. As discussed above that spamming on twitter is different in technique and in nature as compared to other web spams like email spam. Sarita Yardi et al discussed this in a very detailed way in their work [8]. They describe that motivating question for spammers while spamming twitter is that in which way to target and when to target the user. And also what trending topics the spammers should to target and how long they can continue their activities with spamming techniques. Being more practical, Gianluca Stringhini et al [6] explore how the spam has entered in social network sites. They use Random Forest algorithm as a classifier with Weka framework by using features like FF ratio (first feature that compares friend requests that a user sent to the number of friends she has), URL ratio, Message Similarity, Friend Choice and Friend Number. They study how spammers operate to target the social network sites. M. Chuah and M. McCord in [9] discuss some content and user based features as these features are not similar among legal users and spammers.

Zi Chu et al in [10] described that previously all spam detection methods check only individual messages or account for the existence of spam. They focused on the detection of spam campaigns that supervise multiple accounts to spread spam on the twitter network. Alex Hai Wang in [11] proposed a graph model called directed graph model to discover the friend and follower relationship on twitter network. By using Nave Bayesian classifier graph based and content based features are suggested for the detection of spam tweets. In graph based features three features are used namely friends, followers and the reputation of a user is calculated for discovering spam. In content based features duplicate tweets, HTTP links, replies and mentions and trending topics computed for spam detection. In [13], Nikita Spirin studies URLs shared by users on twitter and the estimation of spam for those users who share these links in the network and utilize the information to web spam detection algorithms by proposing a new set of URL derived features for a twitter user representation. Also propose a solution for construction of automatic dataset by analyzing URLs shared by non-spam users in social media for the problem of web spam detection.

In [14] another approach is discussed for spam detection in twitter network. They study the propagation of spam in the network. And they want to find out whether there is a pattern that spammers used for spam proliferation through the network and to determine whether the accounts are either been compromised or overtaken by spammers or certain accounts are purely created for spam activities in the network. They examine the characteristics of the graph of spam tweets and run Trust Rank technique on the collected data. In [15] introduced features for spam tweets detection without earlier statistics of the user and use statistical presentation for the analysis purpose of language to identify spam in twitter topics.

Jonghyuk.S et al in [16] discussed that previously spam detection schemes were based on the features of account information like age of the account, ratio of URLs in tweet and the content similarity of tweet. These features can easily be used by the spammers for spam proliferation activities. They introduced connectivity and distance features (of relation features) for spam detection in twitter which detects spam messages by using connectivity and distance features (of relation features) among the sender of the message and the receiver of the message for checking the spam in the message which is being in progress. Their proposed distance and connectivity features are problematic to operate upon by the spammers and these (relation) features can easily be composed rapidly. Fabricio Benevenuto et al in [12] discussed the problem of detection of spammers in the twitter network as a replacement for spam tweets. The author use social behavior and content based characteristics for the detection of spammers in the twitter network. In [17] spam identification approach is proposed and evaluated for twitter trending topics. Two components of this methodology are detection of timestamp gap among the two consecutive tweets of a user and recognizing the tweet content resemblance amongst the tweets posted by the user.

III. Sentimental and Content-Based Features

We propose sentimental features (combined with content and user based features) as part of our spam detection approach for twitter. All proposed features are described in table **??** in detail.

TABLE II: Features and their Descriptions

Feature	Description
Negative Words Count	Total negative words in a tweet. Negative score computed through SentiWordNet3.0 [18]. It is the sentimental feature.
Negative Words Ratio	It is calculated on the bases of all negative words in a tweet converted in to ratio using equation Negative words Ratio= $\frac{TotalNegativeWords}{TweetLength} \times 100$
Negative Score	Negative Score values are calculated on the bases of sum of all negative words scores of a tweet.
Positive Words Count	Positive Words Count values are calculated on the bases of all positive words in the tweet computed through SentiWordNet3.0 [18]. it is also the sentimental feature.
Positive Words Ratio	Positive Words Ratio value is calculated on the bases of all positive words in a tweet converted in to ratio. Values are calculated on the bases of following formula: Positive words Ratio= $\frac{TotalPositiveWords}{TweetLength} \times 100$

Continued on next column

Continued from previous column

Feature	Description
Positive Score	Positive Score values are calculated on the bases of sum of all positive words scores of a tweet.
Subjectivity Score	Subjectivity Score values are calculated from the tweet on the bases of following formula: Subjectivity Score = Positive Score - Negative Score
Adjectives	This value is calculated on the bases of all adjectives in a tweet with a sentimental value greater than a fixed threshhold. Adjectives are extracted from a Tweet using Part-of-Speech Tagging. It is also a sentimental feature.
Verbs	This value is calculated on the bases of all verbs used in a tweet with a sentimental value greater than a fixed threshhold. Verbs are extracted from Tweet on bases of Part-of-Speech Tagging. It is also a sentimental feature.
Adverbs	This value is calculated on the bases of all adverbs used in a tweet with a sentimental value greater than a fixed threshhold. Adverbs are extracted from Tweet on bases of Part-of-Speech Tagging. It is also a sentimental feature.
Smiles ☺	Smiles values are calculated on the base of all smiles ☺used in a tweet. It is the emotional sentimental feature.
High Smiles (;))	High Smiles values are calculated on the base of all smiles (;)) used in a tweet; High Smiles are extracted from tweet on bases of emotional sentiments (;)); it is also the emotional sentimental feature.
Sad Faces ☹	This is calculated on the base of all sad faces ☹used in a tweet. It is emotional sentimental feature.
Deep Sad Faces	This value is calculated on the base of all ;(used in the tweet text. It is also emotional sentimental feature.
Hashtags Percent	Hashtags percent values are calculated on base list of all Hashtags included in a tweet converted in to percentage. Values are calculated on the bases of following formula: Hashtags Percent= (Total Hashtags)/(Tweet Length) X 100

Continued on next column

Continued from previous column

Feature	Description
URLs Percent	URLs percent values are calculated on the bases of all URLs included in a tweet converted in to percentage. Values are calculated on the bases of following formula: URLs Percent= $\frac{TotalURLs}{TweetLength} \times 100$
Users Mention	It is calculated on the bases of all usernames (@username) mentioned in the tweet text; and converted in percentage. Values are calculated on the bases of following formula: Users Mention Percent= $\frac{TotalUsersMention}{TweetLength} \times 100$

Concluded

IV. Experiments and Evaluations

A. Data Collection

We downloaded tweets for 29 the most trending topics of twitter for year 2012 using APIs provided by twitter. After basic pre-processing, we are left with 29K (1K for each topic) tweets. Manual annotation of these tweets was done with spam or not-spam labels using two annotators A and B. Kappa score [7] for this annotation was found satisfactory (0.82) to proceed with the experiments. We decide to use standard metrics for measuring the usefulness of our approach and hence precision, recall, and F-measure are used.

B. Features Performance Comparison

Here we will discuss our proposed features spam detections performance by using five selected classifiers (SVM, Random Forest, Naive Bayes, Bays Network and J48). We have compared the performance of different features by making different combinations, We have discussing just one combination "all proposed features with baseline features combination" , its performance are given in Table III.

C. All Features and Baseline Features Comparisons

TABLE III: All Features and Baseline Features Accuracy

Classifiers	Baseline Accuracy (%age)	All Features (%age)	Improvement (%age)
BayesNet	90.60	89.76	-0.84
NaiveBayes	14.13	25.30	11.17
Random Forest	91.81	92.29	0.48
J48	91.87	92.34	0.47
LibSVM	91.27	91.41	0.13

Table III shows the accuracy of all features with baseline features by using 10 folds cross validation while figure 1 shows the graphical representation of the information represented in table III. As we have seen in table III result and 1, Naive Bayes spam detections performance improved with our proposed features. Naive Bayes accuracy with baseline features is 14.13%, result improved a lot with our proposed features combination with baseline features to 25.30% (i.e. 11.18% improvement). We have also got good improvement in Random Forest and J48 classifiers. Random Forest with baseline accuracy is 91.81% is improved with all proposed features to 92.29% with gives 0.48% improvements in accuracy while J48 has given 0.47% improvement. SVM has also shown some improvements in spam detection performance (0.14%).

We repeated the experiments using 70% training dataset fetched by using "Remove Percentage Weka"[2] unsupervised filter by setting percentage property to 70% (contain 20141 spam and non-spam tweets) and testing datasets (contain 6042 spam and non-spam tweets) is fetched by setting the "invert selection" properties to false.

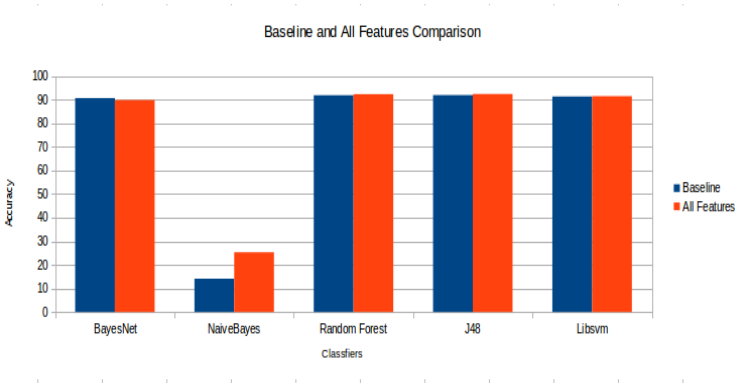

Fig. 1: All Features and Baseline Features Accuracy

Table IV displays results of these experiments while figure 2 shows the graphical representation of all and Baseline Features Accuracy with 70% Training and Testing Datasets (IV).

TABLE IV: All and Baseline Features Accuracy with 70% Training and Testing Datasets

Classifiers	Baseline Accuracy (%age)	All Features (%age)	Improvement (%age)
BayesNet	92.15	91.65	-0.50
NaiveBayes	16.56	26.68	10.12
Random Forest	91.61	92.41	0.80
J48	92.3540	92.20	-0.15
LibSVM	93.37	93.35	-0.02

As we have seen in Figure 2, Naive Bayes and Random Forest spam detections performance improved with our proposed features with 70% training and testing datasets. Naive Bayes accuracy improve further as compare to the previous experiments of 10 fold cross validation (i.e. 25.30% vs 26.68%). Random Forest has also shown some improvements in spam detection performance (0.80%).

[2]http://www.cs.waikato.ac.nz/ml/weka/Weka3

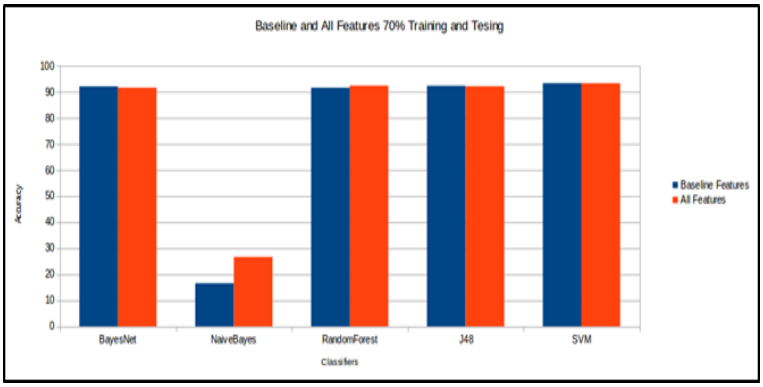

Fig. 2: All Proposed and Baseline Features Accuracy with 70% Training and Testing Datasets

Features Combination with Baseline Features Comparisons Table V shows the accuracy of all combination of features with baseline features by 10 folds using cross validation in percentage values while figure 3 shows its graphical representation.

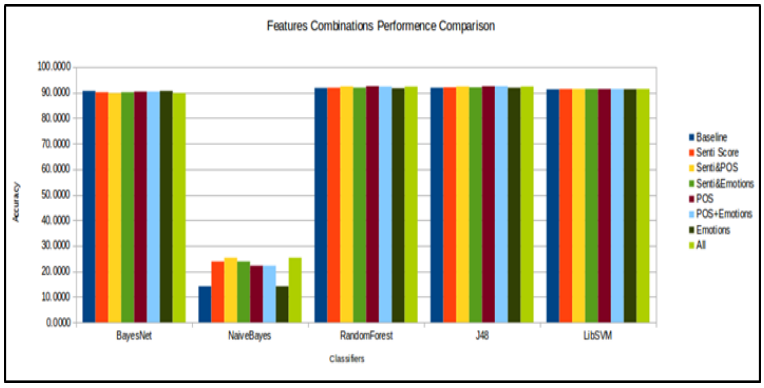

Fig. 3: All Combination with Baseline Features Accuracy

TABLE V: All Combination with Baseline Features Accuracy

Classifiers	Baseline Accuracy	Senti	Senti+POS	Senti+Emotions	POS	POS+Emotions	Emotions	All Combined
BayesNet	90.60	90.15	89.76	90.15	90.37	90.37	90.60	89.76
NaiveBayes	14.13	23.84	**25.30**	23.84	22.16	22.17	14.13	**25.30**
Random Forest	91.81	91.87	92.39	91.90	**92.48**	92.35	91.71	92.29
J48	91.87	92.05	92.36	92.05	**92.46**	**92.46**	91.87	92.34
LibSVM	91.27	91.38	91.40	**91.41**	91.37	91.40	91.32	91.41

As described in the table and figure, for Naive Bayes classifier we have got good improvement in all combinations but the best combination stands "All Combined" while Random Forest gets improvement in "POS sentimental features" combination. With J48 and SVM as we seen we are getting good performance in all features combinations.

At the end, figure 4 shows the percentage accuracy improvement of all combination of features as compared with baseline features by using 10 folds cross validation. The values are calculated by using following formula: Value = Features Combination Accuracy Baseline Features Accuracy

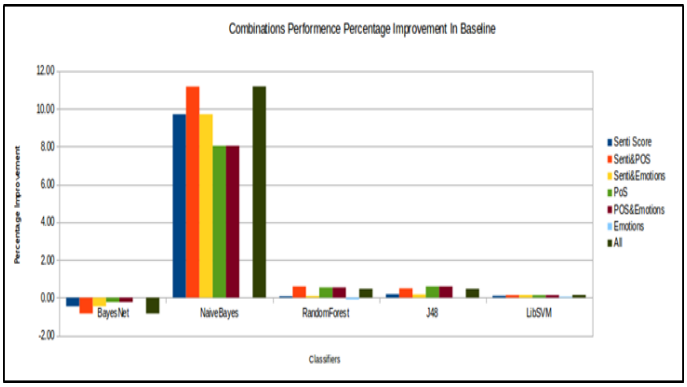

Fig. 4: Combinations Percentage Improvement Compared with Baseline Accuracy

As we have seen in figure 4, Naive Bayes shows good performance as compared with baseline features accuracy. It gained 11.18% improvement as compared with baseline features in all combinations; and with proposed (Sentimental), content and users based features. In Random Forest we have got good percentage performance in Sentimental score and POS features combination with baseline features, its improving 0.59% in spam detection performance with all features combination its just 0.47% performance improvement. In J48 as we have seen its performance improves in POS, POS and emotions combinations with baseline features both have 0.59% improvement with all features combination its just 0.47% performance improvement. SVM also have showing little bit improvement in spam detection accuracy performance its best improvement coming in combination of all proposed features with baseline features gaining 0.14% performance better then as compared with baseline features accuracy. Sentimental score and emotions features combination also have same performance output 0.14%. BayesNet have lost spam detection performance in almost all combinations

V. Conclusion

In this paper, we have suggested some sentimental and POS based features that are combined with content/user based features which can be used to differentiate between spam tweets and legitimate tweets on the twitter a popular online social networking site. Our suggested features are influenced by twitter spam detection policies and our observations of spam behaviors. By using twitter API we collected our dataset of 29 most trending topic in 2012. We proposed sentimental and some content based features which will help in identifying spam tweets and return spam filtered result set when user visit twitter with good accuracy rate. We evaluate the usefulness of our suggested features in spam detection by using five traditional classifiers like BayesNet, Naive Bayes, Random Forest, Support Vector Machine (SVM) and J48 schemes. Our experiments results shows that Naive Bayes, J48 and Random Forest classifier gives over all best performance than the other classifiers like SVM (it shows some improvements in spam detections as compared with content and user based baseline features) and BayesNet. Naive Bayes, Random Forest, J48 and SVM spam detections performance improved with our all proposed features combination. Naive Bayes accuracy with baseline features is 14.1313%, results improved a lot with our proposed features combination with baseline features to 25.3084% and it gives 11.18% performance improvement in spams detections. Random Forest baseline accuracy is 91.8118 % is also improved to 92.2914% which given 0.48% improvement. J48 baseline features accuracy is 91.8778% is improved to 92.3435% which gives 0.47% improvement. SVM baseline features accuracy is 91.2765% with combination to our all proposed features improved to 91.4156% which gives 0.14% performance improvement. By using Naive Bayes, J48 and Random Forest classifier, our suggested features can achieve 93% precision and 95% F-measure. We are leaving future work for now to evaluate our spam detection scheme using larger twitter dataset as well as other online social networking sites like Facebook.

References

[1] A. Einstein, On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat, Annalen der Physik 17, pp. 549-560, 1905.

[2] C. Shirky, Blog explosion and insiders club: Brothers in cluelessness (6 October), at http://many.corante.com/archives/2004/10/06/blog_explosion_and_insiders_club_brothers_in_cluelessness.php, accessed 14 August 2009.

[3] A. Swidler, 2009. Q309 Spam and Virus Trends from Postini (1 October), at http://googleenterprise.blogspot.com/2009/10/q309-spam-virus-trends-from-postini.html,accessed14November2009

[4] Symantec, 2009. State of spam, at http://www.symantec.com/business/theme.jsp?themeid=state_of_spam,accessed14November2009

[5] D. Waters, 2009. Spam overwhelms email messages, BBC News (8 April), at http://news.bbc.co.uk/2/hi/technology/7988579.stm, accessed 4 December 2009.

[6] Gianluca Stringhini, Christopher Kruegel, Giovanni Vigna. Detecting Spammers on Social Networks. In Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC '10). ACM, New York, NY, USA, 1-9

[7] Viera AJ, Garrett JM. Understanding interobserver agreement: the statistic. Fam Med 2005; 37:3603.

[8] Sarita Yardi ,Danial Romero, Grant. S and danah. B Detecting Spam in a twitter Network, First monday, Volume15, Number1-4 January 2010

[9] M. McCord, M. Chuah. Spam Detection on twitter Using Traditional Classifiers 8th International Conference, ATC 2011, Banff, Canada, September 2-4, 2011

[10] Zi Chu, Indra Widjaja, and Haining Wang. Detecting Social Spam Campaigns on twitter. In Proceedings of 10th International Conference, ACNS 2012, Singapore, June 26-29, 2012.

[11] Alex Hai Wang. DONT FOLLOWME: SPAM DETECTION IN TWITTER, Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT), p 1-10

[12] Fabrcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virglio Almeida Detecting Spammers on twitter, Seventh annual Collaboration, Electronic messaging, AntiAbuse and Spam ConferenceJuly 13-14, 2010, Redmond, Washington, US

[13] Nikita Spirin, Mutually Reinforcing Spam Detection on twitter and Web, Technical Report, NorthEastern University

[14] Kanak Biscuitwala, Vidya Ramesh, Kevin Tezlaf Analyzing twitter Spam, Project Report Autumn 2011,Stanford University

[15] Juan Martinez-Romo , Lourdes Araujo. Detecting malicious tweets in trending topics using a statistical analysis, Expert Systems with Applications, Volume 40, Issue 8, 15 June 2013, Pages 2992-3000

[16] Jonghyuk Song, Sangho Lee, and Jong Kim. 2011. Spam filtering in twitter using sender-receiver relationship. In Proceedings of the 14th international conference on Recent Advances in Intrusion Detection (RAID'11), Robin Sommer, Davide Balzarotti, and Gregor Maier (Eds.). Springer-Verlag, Berlin, Heidelberg, 301-317

[17] Puneeta Sharma and SampatBiswas. Identifying Spam in twitter Trending Topics http://www-scf.usc.edu/~sampatbi/pubs/Identifying_Spam_in_twitter_Trending_Topic.pdf

[18] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of LREC 2010