A Report

On

Task 9: Sentiment Analysis in Twitter

Team ID: Indian_Institute_of_Technology-Patna

Team affiliation: Indian Institute of Technology Patna, INDIA

Contact Information: Participant 1. Vikram Singh (vikram.mtcs13@iitp.ac.in)
Participant 2. Mohammed Arif Khan (arif.mtmc13@iitp.ac.in)
Guide: Dr. Asif Ekbal (asif@iitp.ac.in)

Submission: Indian_Institute_of_Technology-Patna.zip

System specification

Core approach

We obtained the Training, Development and Test data sets provided by the organizers, Pre-processed the data, identified and added features which we thought to be relevant in identifying the sentiment of the tweet. We then built up a model using SMO on Training Data, improved the accuracy by checking the model on development data and modifying the features in an iterative manner. Finally we run the model on test data set to obtain the sentiment classification of its tweets.

Supervised or Unsupervised

As the class label was associated with training set so we are following supervised machine learning approach using SMO classifier in which we first built a mod using given training set and then applied it on the given development and test set.

Critical tools used

1. ARK Tagger.
2. Weka 3.6.
3. Eclipse IDE for Java Programming.
4. Notepad++ as text editor.

Critical features used

Features used for sentiment analysis are listed below:

S.N.	Feature Name	Description	Data Type
1	sentiword_positive	Sum of positive polarity of all words of a tweet.	Numeric
2	sentiword_negative	Sum of negative polarity of all words of a tweet.	Numeric
3	sentiword_neutral	Polarity measure for neutral words.	Numeric
4	stop_words	Number of Stop words present in a Tweet.	Numeric
5	all_cap_words	Number of words having all character in upper case for a tweet.	Numeric
6	no_of_hash	Number of Hash tag present in a tweet.	Numeric
7	tweet_length	Number of words in a tweet having more than two characters.	Numeric
8	init_cap	Number of words starting with a Upper Case letter.	Numeric
9	percent_cap	Percentage of capitalized characters in a tweet.	Numeric
10	psmiley	Number of positive smiley present in a tweet.	Numeric
11	nsmiley	Number of negative smiley present in a tweet.	Numeric
12	pwords	Number of positive words present in a tweet.	Numeric
13	nwords	Number of negative words present in a tweet.	Numeric
14	neutralwords	Number of negative words present in a tweet.	Numeric
15	adjective_count	Number of adjectives present in a tweet.	Numeric
16	not_exists	Any form of ‘not’ is present or not.	Boolean
17	repeating_char	Word(s) having consecutive repeated atleast 3 characters is present or not.	Boolean

Significant data pre/post-processing

The following pre-processing was done

1. Tweets were extracted based on markers given (Task A only).
2. Extracted parts were tagged using ARK Tagger.
3. Parts of Tweets which were not related with emotions were filtered out.
4. Feature String was added to every row of tweet.
5. File was converted to .arff format (To be used with Weka).

The following post-processing was done

1. Class predictions generated by Weka were inserted in test file to replace the unknown portion.
2. File was checked to be in correct format using the utility scorer given by organizers.

Other data used (outside of the provided)

1. Smiley list was obtained from Wikipedia.
2. Stop word list was obtained from the internet.

References

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media, LSM '11, pages 30–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

[2] Olena Kummer and Jacques Savoy, Feature Selection in Sentiment Analysis.

[3] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu, NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets.

[4] Hassan Saif, Yulan He and Harith Alani, Semantic Sentiment Analysis of Twitter.

[5] Ahmed Abbasi, Hsinchun Chen, and Arab Salem,Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums.

[6] ChristianRohrdantz, Hewlett Packard Labs, Palo Alto, LARS-ERIK Haug, Daniel A. Keim Feature-based Visual Sentiment Analysis of Text Document Streams.

[7] Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. 2013. SemEval-2013 task 2: Sentiment analysis in twitter. In Proceedings of the International Workshop on Semantic Evaluation, SemEval'13.

[8] Sam Clark and Richard Wicentowski, SwatCS: Combining simple classifiers with estimated accuracy. Second Joint Conference on Lexical and Computational Semantics, Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 425–429, Atlanta, Georgia, June 14-15, 2013.

[9] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.

[10] Apoorv Agarwal, Fadi Biadsy and Kathleen R. Mckeown,Contextual Phrase-Level Polarity Analysis using Lexical Affect Scoring and Syntactic N-grams. Proceedings of the 12th Conference of the European Chapter of the ACL, pages 24–32, Athens, Greece, 30 March – 3 April 2009.

[11] Tim O'Keffe, Irena Koprinska, Feature Selection and Weighted Methods in Sentiment Analysis, Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

[12] Himid Poursepanj, Josh Weissbock, and Diana Inkpen, System description for SemEval 2013 Task 2 Sentiment Analysis in Twitter.