Análisis de sentimientos en Twitter: impacto de las características morfológicas

Miguel Jasso-Hernández[1,2], David Pinto[2], Darnes Vilariño[2], Cupertino Lucero[1]

[1] Universidad Tecnológica de Izúcar de Matamoros, México

[2] Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla, México

migueljhdz18@yahoo.com.mx, {dpinto, darnes}@cs.buap.mx,
cuper_lucero@hotmail.com

Resumen. En este artículo se presentan una serie de experimentos encaminados al análisis de sentimientos sobre textos escritos en Twitter. En particular, se estudian diversas características morfológicas para la representación de los textos con la finalidad de determinar aquellas que proporcionan el mejor rendimiento en el momento de detectar la carga emocional contenida en los Tweets.

Palabras clave: Análisis de sentimientos, etiquetas morfológicas, Tweets.

1. Introducción

Analizar la carga emocional en textos es una tarea que reviste una gran importancia en la actualidad. Existe una multitud de aplicaciones que pueden resultar beneficiadas de procedimientos computacionales que permitan detectar, automáticamente, si la intención del autor ha sido expresarse de manera "positiva", "negativa", "objetiva" o "neutral". Consideremos, por ejemplo, el caso de una personalidad política que requiere saber si la comunidad tiene una apreciación positiva o negativa sobre su persona. Otro ejemplo, podría ser la determinación de la reputación para una institución pública o privada. En cualquiera de los dos ejemplos, existe la necesidad de analizar el punto de vista de las personas (usuarios) sobre nuestras entidades destino. Si bien, en el pasado era común aplicar cuestionarios hacia los usuarios, en la actualidad esta práctica es raramente usada debido principalmente a los siguientes inconvenientes:

1. La aplicación de cuestionarios es un proceso costoso, desde el punto de vista del tiempo y económico.
2. El concentrado de dichos cuestionarios requiere tiempo y un análisis posterior.
3. La selección de candidatos sobre los cuales se aplica el cuestionario debe ser decidido con cuidado para garantizar que los resultados de análisis sean apropiados (cantidad y calidad).
4. El análisis de los datos tiene que hacerse con prontitud para evitar que las conclusiones sean obsoletas.

De esta manera, es mucho más práctico y conveniente usar datos frescos obtenidos directamente de las redes sociales. Las personas suelen expresarse libremente sobre los temas que son de su interés. El único problema es que dichos datos son expresados en lenguaje natural, y por tanto requieren de métodos computacionales automáticos para su tratamiento. Aun así, esta aproximación resulta ser mucho más atractiva para las empresas y ha resultado en una área de investigación sumamente activa dentro de la comunidad relacionada con el procesamiento automático del lenguaje natural.

El objetivo de este trabajo es el de evaluar el impacto en el uso de diversas características morfológicas sobre la tarea de la detección de carga emocional en Tweets (*positive*, *negative*, *neutral* y *objective*). El problema se ha planteado solucionar desde la perspectiva de la clasificación de textos usando métodos de aprendizaje automático. Esta perspectiva se ha tomado en cuenta basado en la existencia de un corpus supervisado (con clasificación manual).

A continuación se presentan algunos trabajos relacionados con el análisis automático de sentimientos en textos obtenidos desde las redes sociales.

1.1. Trabajos relacionados

Existen estudios relacionados con la identificación de emociones en Twitter, sin embargo, pocos de ellos prestan atención al análisis de la aportación parcial que generan las características morfológicas. Por ejemplo, en [1] se calcula la probabilidad de polaridad a priori asociada con etiquetas de partes de la oración (PoS). Se usan hasta 100 características adicionales que incluyen el uso de emoticones y diccionarios de palabras positivas y negativas. Los resultados reportados muestran hasta un 60 % de exactitud. Por otro lado, en [13] se propone una estrategia que hace uso de pocos recursos léxicos; en particular, utiliza relaciones discursivas tales como conectividad y condicionals para incorporarlas en los modelos clásicos de bolsas de palabras con la intención de mejorar los valores de exactitud sobre el proceso de clasificación. También se prueba la influencia de operadores semánticos tales como los modales y las negaciones y el grado en que afectan el sentimiento presente en una oración.

Uno de los mayores avances obtenidos en la tarea de análisis de sentimientos ha sido en un competencia propuesta en el marco de SemEval 2013 [3,4,5,6,7,8,9, 10,11,12,13,14,15]. Algunos de los trabajos han permitido tener un panorama amplio de diversos métodos y características usadas en la tarea mencionada. No cabe duda que esta es una tarea importante que será de atención por la comunidad de lingüística computacional en los años siguientes.

2. Experimentos

En esta sección se describen los experimentos llevados a cabo. A continuación se describe el conjunto de datos usado; posteriormente, las características evaluadas para representar a los textos; enseguida el tipo de clasificador usado; para finalmente mostrar los resultados obtenidos.

2.1. Conjunto de datos usado

En nuestros experimentos hemos hecho uso de un conjunto de datos de entrenamiento y prueba proporcionados en el marco de la competencia SemEval 2014; en particular, en la Tarea 9 que ha sido denominada "Análisis de Sentimientos en Twitter"[1]. Este corpus presenta 6,162 tweets escritos en el idioma inglés, los cuales han sido etiquetados manualmente con las siguientes clases: *positive*, *negative*, *neutral* y *objective*, los cuales pueden ser utilizados como datos de entrenamiento. El conjunto de textos de prueba contiene 978 tweets. Una descripción de sus características generales puede ser vista en la Tabla 1.

Tabla 1. Características del corpus de evaluación.

Característica	**Corpus de entrenamiento**	**Corpus de prueba**
No. de Tweets	6,162	978
No. de Palabras	100,973	15,992
Vocabulario	17,000	4,611
Longitud promedio	16.38	16.35
Tweets positivos	2,252	356
Tweets negativos	874	184
Tweets neutrales	1,062	330
Tweets objetivos	1,974	108

El vocabulario del corpus de prueba comparte 3,380 términos con el vocabulario del corpus de entrenamiento. Lo cual significa que el 73 % de su vocabulario está presente en el vocabulario usado en el entrenamiento. Sin embargo, dado el tamaño del vocabulario de entrenamiento (17,000) podemos ver que solamente un 15 % del vocabulario es común entre ambos corpora, lo cual muestra claramente que existe una gran cantidad de palabras que no serán útiles en la tarea de clasificación.

2.2. Descripción de las características usadas

Tal y como se ha mencionado, el objetivo de este trabajo consiste en evaluar el impacto que pueden tener las características morfológicas sobre el proceso de representación textual, cuando se lleva a cabo una tarea de identificación de carga emocional en Tweets. Así, hemos filtrado cada palabra de los Tweets del corpus de entrenamiento y prueba, dejando únicamente aquellas palabras que cumplan con una y solamente una etiqueta morfológica. El proceso de etiquetado se llevó a cabo usando el etiquetador TreeTagger[2]. Las etiquetas usadas pueden ser vistas en la Tabla 2.

Dado que los Tweets son textos sumamente cortos, en algunos casos resulta que la cantidad de palabras que cumple con alguna etiqueta PoS es cero, por

[1] `http://alt.qcri.org/semeval2014/task9/index.php?id=data-and-tools`
[2] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

Tabla 2. Etiquetas morfológicas usadas en el proceso de clasificación.

Etiqueta PoS	Descripción
JJ	Adjetivo
NN	Sustantivo en singular
VBN	Verbo en pasado participio
VB	Verbo en su forma base
RB	Adverbio
IN	Intersección
NP	Nombre propio en singular
PP	Preposición
RBR	Adverbio comparativo
RBS	Adverbio superlativo
RP	Partícula
VBG	Verbo en gerundio o pasado participio
JJR	Adjetivo comparativo
JJS	Adjetivo superlativo
MD	Modal
NPS	Nombre propio en plural
PDT	Predeterminante
VBZ	Verbo en presente, tercera persona singular
VBP	Verbo en presente, no tercera persona singular
WDT	Determinante tipo Wh
WP	Pronombre tipo Wh
WPS	Pronombre posesivo tipo Wh
WRB	Adverbio tipo Wh
NNS	Sustantivo en plural

tanto se tomó la decisión de seleccionar las primeras cinco palabras en aquellos casos en los cuales la selección de palabras fuese nula. Esto permite homogeneizar los resultados haciendolos comparables a través de todas las etiquetas morfológicas, sobre todo pensando en que a corto plazo nos gustaría realizar un ensamble de los resultados por categoría morfológica. En este caso, se necesitaría que cada instancia tenga, al menos, una característica.

2.3. Clasificador usado

Se seleccionaron tres clasificadores de distinta naturaleza con la finalidad de evaluar el proceso adecuadamente. En particular, se usó un clasificador basado en árboles de decisión conocido como J48, el cual es una implementación del algoritmo C4.5, uno de los algoritmos de minería de datos más utilizado. También se uso una implementación del algoritmo de Naïve Bayes, el cual calcula la probabilidad de cada característica, dada la característica anterior. Finalmente, se empleó una implementación de las máquinas de soporte vectorial (SVM) conocida como SMO. Una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta que puede ser utilizado en problemas

de clasificación o regresión. El objetivo es buscar una buena separación entre las clases, lo cual permitirá un clasificación correcta.

Se usaron las implementaciones de J48, Naive Bayes y SMO presentes en la herramienta WEKA[3], empleando los parámetros default existentes en cada clasificador.

2.4. Resultados

En esta sección se presentan los resultados obtenidos sobre cada uno de los corpora (entrenamiento y prueba), usando las características comentadas anteriormente. Es importante mencionar que los experimentos se realizaron sobre cada una de las etiquetas PoS de forma independiente. La Tabla 3 muestra el porcentaje de datos que fueron clasificados correctamente (C) y el porcentaje de aquellos que fueron clasificados incorrectamente (I). Hemos resaltado los valores más altos, los cuales, son obtenidos principalmente por el clasificador basado en máquinas de soporte vectorial (SMO).

En la Figura 1 se observa que el comportamiento obtenido por SMO es superior, sin importar, el tipo de etiqueta morfológica utilizada. En particular, el orden de importancia en el proceso de clasificación es el siguiente: JJ, PP, RB, VB, VBP, NN, IN, JJS, RBR, RBS, WP, NPS, VBG, WRB, WPS, PDT, JJR, VBN, WDT, NNS, RP, NP, VBZ, MD. Es decir, de acuerdo a los resultados obtenidos en el corpus de entrenamiento se observa que las cinco características de mayor peso son: adjetivo, preposición, adverbio, verbo en su forma base y verbo en presente (no tercera persona singular).

Para el caso del corpus de prueba (ver Figura 2), se observa que existen únicamente dos casos en los cuales el algoritmo de clasificación SMO no obtiene los mejores resultados (preposición y adverbio). En estos dos casos, el algoritmo de Naïve Bayes obtiene mejores resultados que SMO. El orden de importancia de cada etiqueta morfológica varia bastante con respecto al obtenido sobre el corpus de entrenamiento. Este orden es el siguiente: NN, JJ, VBP, RB, NP, JJS, WRB, VBN, WP, MD, VBZ, VB, PDT, JJR, NNS, RBS, WPS, RP, RBR, VBG, WDT, IN, NPS, PP. Es decir, de acuerdo a los resultados obtenidos en el corpus de prueba se observa que las cinco características de mayor peso son: sustantivo en singular, adjetivo, verbo en presente (no tercera persona singular), adverbio y nombre propio en singular.

De esta manera, podemos observar que las siguientes características son las más importantes en ambos corpora (entrenamiento y prueba): adjetivo, verbo en presente (no tercera persona singular) y adverbio. Es curioso que la preposición se encuentre como una característica importante en el corpus de entrenamiento, mientras que en el corpus de prueba resultó ser la característica de menor importancia.

Estos resultados podrían ser mejorados al buscar un ensamble de características morfológicas que mejor representen a los Tweets. Se deberá hacer una análisis exhaustivo para evitar considerar características que dupliquen su grado

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

Tabla 3. Resultados del proceso de clasificación usando solamente las características morfológicas.

	J48				Naïve Bayes				SMO			
Etiqueta	Entrenamiento		Prueba		Entrenamiento		Prueba		Entrenamiento		Prueba	
PoS	C	I	C	I	C	I	C	I	C	I	C	I
JJ	0.448	0.552	0.305	0.695	0.432	0.568	0.332	0.668	**0.467**	0.533	**0.340**	0.660
NN	0.430	0.570	0.308	0.692	0.421	0.579	0.323	0.677	**0.452**	0.548	**0.343**	0.657
VBN	0.434	0.566	0.314	0.686	0.412	0.588	0.313	0.687	**0.443**	0.557	**0.325**	0.675
VB	0.426	0.574	0.300	0.700	0.407	0.593	0.318	0.682	**0.455**	0.545	**0.322**	0.678
RB	0.445	0.555	0.318	0.682	0.424	0.576	**0.340**	0.660	**0.460**	0.540	0.337	0.663
IN	0.416	0.584	0.309	0.691	0.407	0.593	0.316	0.684	**0.452**	0.548	**0.318**	0.682
NP	0.423	0.577	0.316	0.684	0.413	0.587	0.305	0.695	**0.441**	0.559	**0.332**	0.668
PP	0.437	0.563	0.312	0.688	0.415	0.585	**0.316**	0.684	**0.464**	0.536	0.312	0.688
RBR	0.433	0.567	0.317	0.683	0.414	0.586	0.310	0.690	**0.448**	0.552	**0.319**	0.681
RBS	0.439	0.561	0.308	0.692	0.414	0.586	0.310	0.690	**0.446**	0.554	**0.321**	0.679
RP	0.428	0.572	0.300	0.700	0.404	0.596	0.317	0.683	**0.442**	0.558	**0.321**	0.679
VBG	0.432	0.568	0.315	0.685	0.412	0.588	0.305	0.695	**0.444**	0.556	**0.319**	0.681
JJR	0.436	0.564	0.313	0.687	0.414	0.586	0.310	0.690	**0.443**	0.557	**0.322**	0.678
JJS	0.442	0.558	0.305	0.695	0.415	0.585	0.312	0.688	**0.451**	0.549	**0.329**	0.671
MD	0.426	0.574	0.294	0.706	0.413	0.587	0.320	0.680	**0.440**	0.560	**0.324**	0.676
NPS	0.434	0.566	0.308	0.692	0.412	0.588	0.309	0.691	**0.445**	0.555	**0.316**	0.684
PDT	0.435	0.565	0.310	0.690	0.413	0.587	0.309	0.691	**0.443**	0.557	**0.322**	0.678
VBZ	0.435	0.565	0.300	0.700	0.414	0.586	0.322	0.678	**0.441**	0.559	**0.323**	0.677
VBP	0.435	0.565	0.302	0.698	0.417	0.583	0.328	0.672	**0.452**	0.548	**0.338**	0.662
WDT	0.429	0.571	0.309	0.691	0.413	0.587	0.312	0.688	**0.442**	0.558	**0.319**	0.681
WP	0.434	0.566	0.309	0.691	0.413	0.587	0.310	0.690	**0.445**	0.555	**0.324**	0.676
WPS	0.434	0.566	0.308	0.692	0.412	0.588	0.309	0.691	**0.444**	0.556	**0.321**	0.679
WRB	0.432	0.568	0.306	0.694	0.411	0.589	0.310	0.690	**0.444**	0.556	**0.325**	0.675
NNS	0.433	0.567	0.315	0.685	0.414	0.586	0.315	0.685	**0.442**	0.558	**0.322**	0.678

de representatividad y buscar solamente aquellas que en su conjunto representen mejor a todos los datos.

3. Conclusiones

En este trabajo se analizaron diversas características morfológicas con la finalidad de determinar el grado de importancia de cada una de ellas en la tarea de análisis de sentimientos en Twitter. Se usaron tres diferentes clasificadores sobre datos de entrenamiento y prueba, observando que las características estables a seleccionar en esta tarea deberían ser: adjetivo, verbo en presente (no tercera persona singular) y adverbio.

Como trabajo inmediato, nos proponemos investigar si el ensamble de diversas características podría ayudar a mejorar los resultados obtenidos. Es decir, nos gustaría investigar si ciertas características morfológicas son mutuamente excluyentes, o al menos bastante excluyentes, para poder representar los textos como una mezcla de características morfológicas.

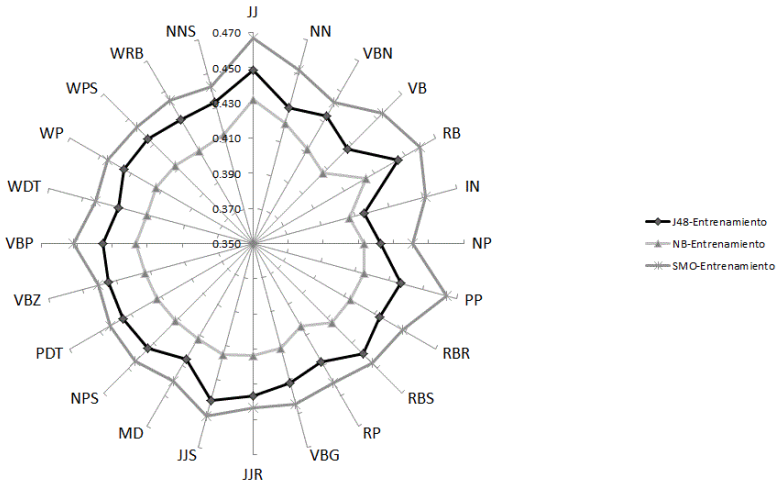

Fig. 1. Resultados obtenidos para cada etiqueta morfológica sobre el corpus de entrenamiento.

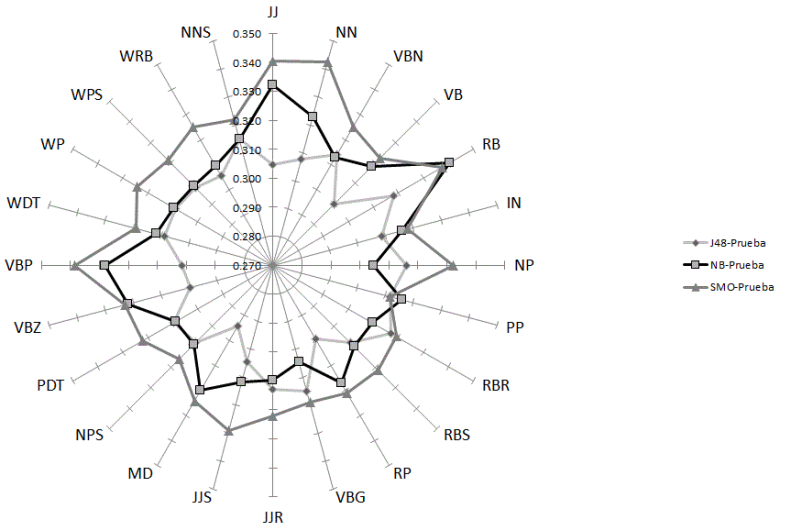

Fig. 2. Resultados obtenidos para cada etiqueta morfológica sobre el corpus de prueba.

Un aspecto interesante sería el de observar cómo se comporta un clasificador con las tres etiquetas que parecen ser las mejores para el conjunto de entrenamiento y prueba, así como las cinco etiquetas morfológicas que obtuvieron los mejores resultados para el conjunto de entrenamiento, por un lado, y por otro, las cinco que se comportaron mejor para el conjunto de prueba. En un trabajo siguiente nos planteamos hacer este análisis.

En nuestros experimentos hemos usado TreeTagger para obtener las categorías morfológicas, sin embargo, este etiquetador no está optimizado para etiquetar textos tipo Tweet, por tanto, sería conveniente hacer pruebas con un etiquetador PoS especial para Twitter como el de CMU[4].

Referencias

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Language in Social Media (LSM 2011). pp. 30–38. Association for Computational Linguistics, Portland, Oregon (June 2011)
2. Balage Filho, P., Pardo, T.: Nilc_usp: A hybrid system for sentiment analysis in twitter messages. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 568–572. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)
3. Balahur, A., Turchi, M.: Improving sentiment analysis in twitter using multilingual machine translated data. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. pp. 49–55. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria (September 2013)
4. Becker, L., Erhart, G., Skiba, D., Matula, V.: Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 333–340. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)
5. Chawla, K., Ramteke, A., Bhattacharyya, P.: Iitb-sentiment-analysts: Participation in sentiment analysis in twitter semeval 2013 task. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 495–500. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)
6. Clark, S., Wicentwoski, R.: Swatcs: Combining simple classifiers with estimated accuracy. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 425–429. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)
7. Hamdan, H., Béchet, F., Bellot, P.: Experiments with dbpedia, wordnet and sentiwordnet as resources for sentiment analysis in micro-blogging. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 455–459. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)
8. Han, Q., Guo, J., Schuetze, H.: Codex: Combining an svm classifier and character n-gram language models for sentiment analysis on twitter text. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 520–524. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)

[4] `http://www.ark.cs.cmu.edu/TweetNLP/`

9. Levallois, C.: Umigon: sentiment analysis for tweets based on terms lists and heuristics. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 414–417. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)
10. Marchand, M., Ginsca, A., Besançon, R., Mesnard, O.: [lvic-limsi]: Using syntactic features and multi-polarity words for sentiment analysis in twitter. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 418–424. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)
11. Martínez-Cámara, E., Montejo-Ráez, A., Martín-Valdivia, M.T., Ureña López, L.A.: Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 402–407. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)
12. Moreira, S., Filgueiras, J.a., Martins, B., Couto, F., Silva, M.J.: Reaction: A naive machine learning approach for sentiment classification. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 490–494. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)
13. Mukherjee, S., Bhattacharyya, P.: Sentiment analysis in Twitter with lightweight discourse analysis. In: Proceedings of COLING 2012. pp. 1847–1864. The COLING 2012 Organizing Committee, Mumbai, India (December 2012)
14. Reckman, H., Baird, C., Crawford, J., Crowell, R., Micciulla, L., Sethi, S., Veress, F.: teragram: Rule-based detection of sentiment phrases using sas sentiment analysis. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 513–519. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)
15. Tiantian, Z., Fangxi, Z., Lan, M.: Ecnucs: A surface information based system description of sentiment analysis in twitter in the semeval-2013 (task 2). In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 408–413. Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013)