

# Detección de posturas en los tweets de las redes sociales mediante la fusión de listas clasificadas y sentimientos

Presentación realizada por:

- Ariel Bravo
- Patricio Fajardo
- Vanessa Romero

# Contenido

- Introducción
- Antecedentes
- Metodologías
- Resultados
- Conclusiones
- Referencias

Information Fusion 67 (2021) 29–40

Contents lists available at ScienceDirect

Information Fusion

journal homepage: [www.elsevier.com/locate/inffus](http://www.elsevier.com/locate/inffus)

Full length article

A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments

Abdulrahman I. Al-Ghadir <sup>a</sup>, Aqil M. Azmi <sup>a,\*</sup>, Amir Hussain <sup>b</sup>

<sup>a</sup> Department of Computer Science, College of Computer & Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

<sup>b</sup> School of Computing, Edinburgh Napier University, Edinburgh EH10 5DT, UK

---

ARTICLE INFO

Keywords:  
stance detection  
Sentiment analysis  
Top-k  
K-NN variants  
Support vector machines  
Twitter

ABSTRACT

Stance detection is a relatively new concept in data mining that aims to assign a stance label (favor, against, or none) to a social media post towards a specific pre-determined target. These targets may not be referred to in the post, and may not be the target of opinion in the post. In this paper, we propose a novel enhanced method for identifying the writer's stance of a given tweet. This comprises a three-phase process for stance detection: (a) tweets preprocessing; here we clean and normalize tweets (e.g., remove stop-words) to generate words and stems lists, (b) feature generation; in this step, we extract and fuse two dictionaries for generating features (top-k and K-NN), (c) classification; in this step, we classify the tweets based on the stance identified by the features generated in the previous step. Our innovative feature selection proposes fusion of two ranked lists (top-k) of term frequency-inverse document frequency (*tf-idf*) scores and the sentiment information. We evaluate our method using six different classifiers: *K* nearest neighbor (*K*-NN), discernibility-based *K*-NN, weighted *K*-NN, class-based *K*-NN, exemplar-based *K*-NN, and Support Vector Machines. Furthermore, we investigate the use of Principal Component Analysis and study its effect on performance. The model is evaluated on the benchmark dataset (SemEval-2016 task 6), and the results significance is determined using *t*-test. We achieve our best performance of macro *F*-score (averaged across all topics) of 76.45% using the weighted *K*-NN classifier. This tops the current state-of-the-art score of 74.44% on the same dataset.

---

1. Introduction

Stance detection is an automatic process of determining whenever the author is likely in favor, against, or neutral towards a proposition

are other applications that may benefit from the automatic stance detection, including information retrieval, text summarization, opinion summarization [7], rumor verification [8], etc.

We can formulate the task of stance detection as follows: given a

# Introducción

Con el auge de la redes sociales (facebook, instagram, twitter, etc) en la actualidad en análisis de los datos es de utilidad para la toma de decisiones.

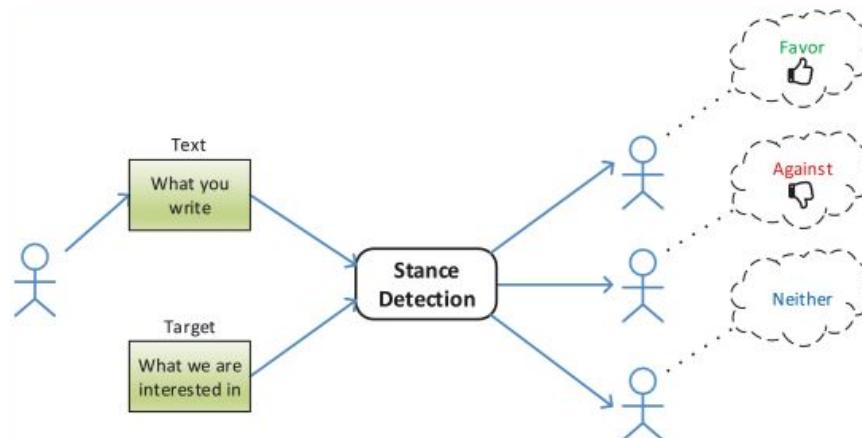
La detección de posturas es un concepto nuevo en la minería de datos que tiene como objetivo asignar una etiqueta de postura (a favor, en contra o ninguna ) a publicaciones en dichas redes sociales.



# Introducción

## ¿Qué es la detección de postura?

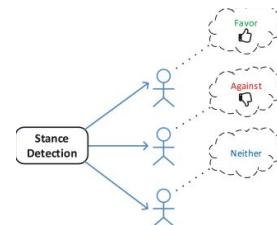
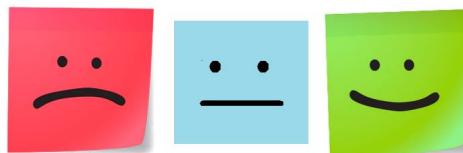
“Es un proceso automático para determinar cuando es probable que el autor esté a favor, en contra o neutral hacia un proposición u objetivo en el texto”. El objetivo puede ser una persona, una organización, un partido político, entre otros.



# Introducción

¿Cuál es la diferencia entre la detección de postura y el análisis de sentimiento.?

- El análisis de sentimiento es un aspecto importante en el detección de posturas.
- Ambos son subcampos de PNL
- Ambos comparten connotaciones similares y usan algunas metodologías de PNL
- En el análisis de sentimientos interesa saber si un fragmento de texto es positivo, negativo o neutral basado solo en el lenguaje utilizado.
- La detección de postura, por otro lado, se define para un tema de destino y puede ser independiente del sentimiento del lenguaje utilizado.



# Introducción

## ¿Campo que se utiliza la detección de posturas?

1. Principalmente se ha utilizado en el ámbito de los debates políticos e ideológicos
2. También ha sido un paso previo al procesamiento en la lucha contra las noticias falsas.
3. Recuperacion de informacion
4. Resumen de texto
5. Resumen de opiniones
6. Verificación de rumores, etc

# Introducción

## Ejemplo de un detección de postura:

Se da un texto de tweet y una entidad objetivo (por ejemplo, una persona u organización), determine si el autor del tweet está a favor, en contra o indiferente al objetivo dado

Objetivo	Tweet	Postura
Legalización del aborto	¿Por qué las bacterias se consideran vida en Marte, pero un latido no se considera vida en la Tierra? #heartbeat #SemST	En contra
El cambio climático es una preocupación	Necesitamos trabajar con confianza, transparencia y guiados por el consenso @manupulgarvidal en el evento @UN_PGA en # action2015 #SemST	Ninguna
Hillary Clinton	Si no estás viendo el discurso de @ HillaryClinton en este momento, te estás perdiendo sus toneladas de sabiduría. #SemST	Favor

# Introducción

## Propuesta del Paper

Proponen un método mejorado para identificar la postura del escritor de un tweet determinado. El proceso que realiza es el siguiente:

- Preprocesamiento de tweets
- Generación de características
- Clasificación

# Trabajos relacionados

En la revisión de trabajos relacionados lo divide en dos partes:

- Detección de posturas que se basaron artesanalmente en la extracción de características.
- Extracción de características (basada en Deep Learning)

# Dataset

El dataset cuenta con tweets de 5 temas objetivo:

- Ateísmo
- El cambio climático es una preocupación real
- Movimiento feminista
- Hillary Clinton
- Legalización del aborto

Los datos consisten en:

- Tema
- Tweet
- Postura

Etiquetas

- A favor
- En contra
- Neutral

# Metodología

Proceso para la detección de posturas

1. Preprocesamiento de tweets
2. Generación de características
3. Clasificación

# 1. Preprocesamiento de tweets

- Se elimina cualquier carácter que no esté en el rango alfabético del inglés.
- Se normaliza la secuencia de caracteres repetidos

Holaaaa → Hola

- Se eliminan las stop-words y se hace una corrección ortográfica
- Generar una lista de palabras y de raíces mediante tokenización

## 2. Generación de características

### Generación de diccionarios

- Revisar cada tweet en el conjunto de datos de entrenamiento.
- Crear los diccionarios Dw y Ds.
- Los diccionarios se utilizan para calcular los valores tf-idf

## 2. Generación de características

### Generación del vector de características

El vector de características consta de tres categorías:

- top-k word list (SVw)
- top-k stems list (SVs)
- Atributo de sentimiento

### 3. Clasificación

Para evaluar el conjunto de características se utiliza:

- SVM
- K-NN

Variantes:

- Discernibility-based K-NN (DKNN)
- Weighted K-NN (WKNN)
- Class-based K-NN (CKNN)
- Exemplar-based K-NN (EKNN)

# Evaluación y Resultados

## Métricas de Evaluación

- Se utiliza el F score
- La métrica oficial aprobada por los organizadores de la tarea de detección de posturas
- Todos los modelos se evaluaron usando únicamente el macro-promedio de dos etiquetas
- Es decir, la media de la puntuación  $F$  para las dos clases principales "a favor" y "en contra"

$$F_{\text{avg}} = \frac{F_{\text{favor}} + F_{\text{against}}}{2}$$

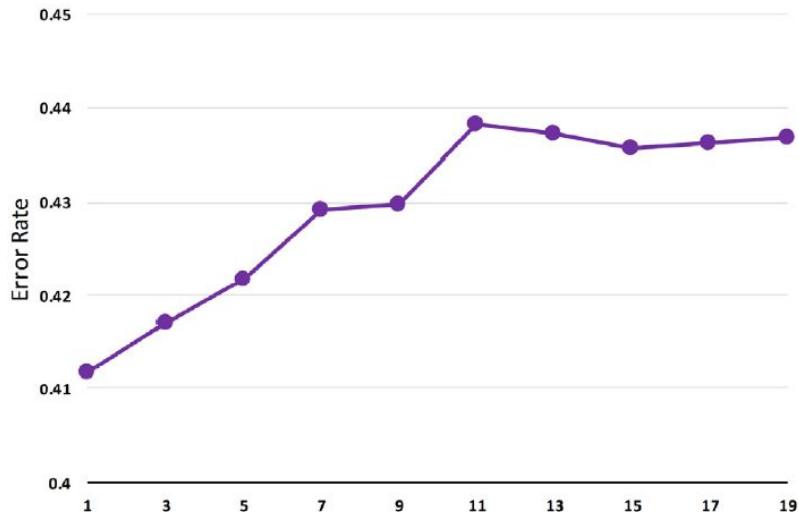
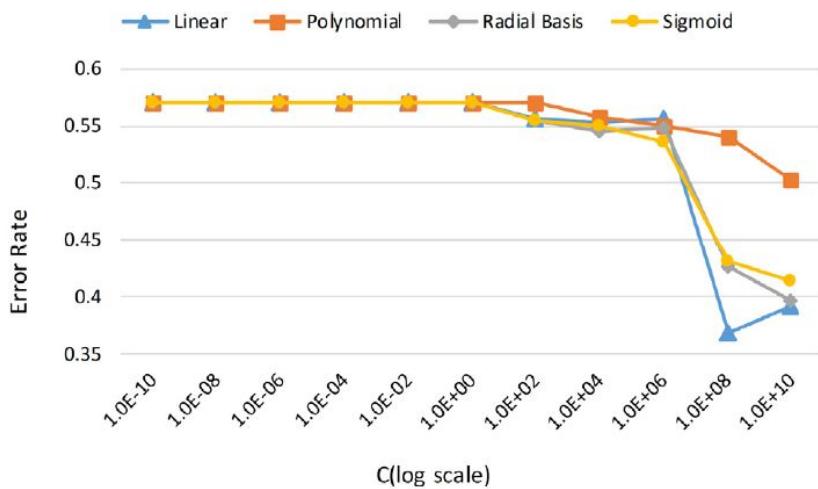
# Evaluación y Resultados

## Tuneo de Parámetros

- Para realizar los experimentos, se realiza un tuneo de ciertos parámetros
- Para el primer experimento con SVM se realiza un tuneo del parámetro C con diferentes funciones de kernels.
- Para el segundo experimento se intenta encontrar el mejor valor de  $K$  para el clasificador K-NN
- Se evalúan los diferentes experimentos usando five-fold cross validation en los datos de entrenamiento.

# Evaluación y Resultados

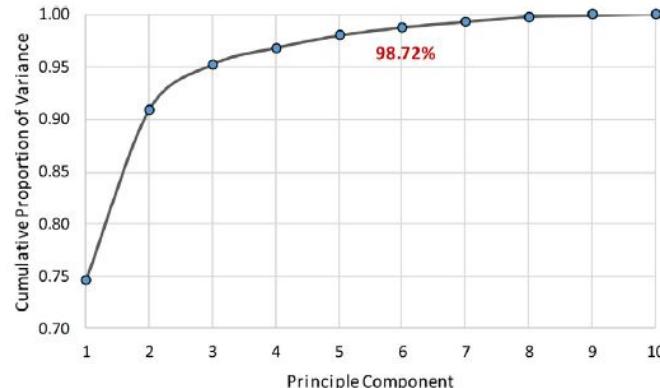
## Tunear de Parámetros



# Evaluación y Resultados

## Tuneo de Parámetros

- También se realizó una reducción de dimensiones aplicando PCA
- De las 10 dimensiones iniciales(debido a un ranking de lista de palabras), después de aplicar PCA se redujo a 6 dimensiones



# Evaluación y Resultados

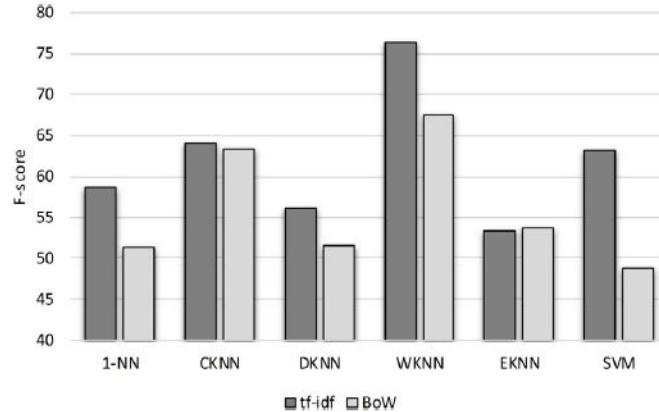
## Resultados experimentales

	$F_{avg}$	
	Overall	
	w/Sent	w/o Sen
1-NN	61.81	58.65
1-NN+PCA	61.49	58.34
CKNN	71.57	64.12
CKNN+PCA	75.19	75.50
DKNN	54.96	56.15
DKNN+PCA	54.47	55.57
WKNN	76.00	76.45
WKNN+PCA	76.26	76.20
EKNN	56.65	53.27
EKNN+PCA	62.03	58.82
SVM	55.37	63.13
SVM+PCA	60.19	65.26

# Evaluación y Resultados

## Resultados experimentales

- Se realiza un experimento paralelo con una diferencia en la creación del vector de características
- El vector de características actualmente analizado utiliza la puntuación tf-idf de las palabras y stems(raíces) top- $k$ .
- Se compara el desempeño cuando tf-idf se reemplaza con Bag-of-Words como score para el vector de características



# Evaluación y Resultados

## Discusión

- Se utiliza la prueba t para determinar si el mejor modelo obtenido anteriormente es estadísticamente significativo en comparación con los otros modelos.

WKNN	p-value	Statistically significant?
w/sentiment vs without	9.13E-01	No
w/PCA vs without	9.31E-01	No
Using <i>tf-idf</i> vs BoW	3.61E-02	Yes
vs 1-NN	1.04E-05	Yes
vs EKNN	2.37E-08	Yes
vs CKNN	1.58E-04	Yes
vs DKNN	7.18E-07	Yes
vs SVM	1.00E-04	Yes

# Conclusiones

- Las redes sociales son parte de la rutina diaria de la mayoría de la población mundial. El objetivo de este estudio es detectar la postura de los tweets, es decir determinar si el autor del tweet está a favor de un objetivo determinado, en contra de él o tiene una postura neutral hacia él.
- Se evaluó el enfoque utilizando diferentes conjuntos de clasificadores, que incluían SVM y diferentes variantes de  $K$  vecino más cercano (K-NN).
- El mejor rendimiento de la  $F$ -score, es de 76,45% es cuando se utiliza el clasificador WKNN, superando el rendimiento del estado del arte en un 72,01%.
- Se demostró que, de hecho, este es un resultado estadísticamente significativo independientemente de si se incluyó o no la información de sentimiento, y si se aplicó PCA o no.
- El método propuesto funciona bien con documentos bastante cortos con contexto limitado, como tweets.

# Referencia

A. Al-Ghadir, A. Azmi, y A. Hussain, «A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments», *Information Fusion*, vol. 67, pp. 29-40, oct. 2020, doi: [10.1016/j.inffus.2020.10.003](https://doi.org/10.1016/j.inffus.2020.10.003).